Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 33

11-1-2009

Vol. 8, No. 2 (Full Issue)

JMASM Editors

Journal of Modern Applied Statistical Methods
November, 2009, Vol. 8, No. 2, 355-669

1538 – 9472/09/$95.00

Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
November, 2009, Vol. 8, No. 2, 355-669

1538 – 9472/09/$95.00

Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
November, 2009, Vol. 8, No. 2, 355-669

1538 – 9472/09/$95.00

Journal of Modern Applied Statistical Methods
November, 2009, Vol. 8, No. 2, 355-669

1538 – 9472/09/$95.00

JMASM is an independent print and electronic journal (http://www.jmasm.com/), publishing (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal of Modern Applied Statistical Methods
November, 2009, Vol. 8, No. 2, 355-359

1538 – 9472/09/$95.00

INVITED ARTICLE

Analysis of MultiFactor Experimental Designs

Phillip Good
Information Research
Huntington Beach, CA.

In the one-factor case, Good and Lunneborg (2006) showed that the permutation test is superior to the analysis of variance. In the multi-factor case, simulations reveal the reverse is true. The analysis of variance is remarkably robust against departures from normality including instances in which data is drawn from mixtures of normal distributions or from Weibull distributions. The traditional permutation test based on all rearrangements of the data labels is not exact and is more powerful that the analysis of variance only for 2xC designs or when there is only a single significant effect. Permutation tests restricted to synchronized permutations are exact, but lack power.

Key words: analysis of variance, permutation tests, synchronized permutations, exact tests, robust tests, two-way experimental designs.

Introduction

Tests of hypotheses in a multifactor analysis of variance (ANOVA) are not independent of one another and may not be most powerful. These tests are derived in two steps: First, the between-cell sum of squares is resolved into orthogonal components. Next, to obtain p- values, the orthogonal components are divided by the within-cell sum of squares. As they share a common denominator, the test statistics of main effects and interactions are *not* independent of one another. On the plus side, Jagers (1980) showed that if the residual errors in the linear model are independent and identically distributed, then the distribution of the resultant ratios is closely approximated by an F-distribution even if the residual errors are *not* normally distributed. As a result, ANOVA p-values are almost exact.

But are ANOVA tests the most powerful? In the one-way design (the one-factor case), Good and Lunneborg (2005) found that tests whose p-values are based on the permutation distribution of the F-statistic rather than the F-distribution are both exact and more

Phillip Good is a statistical consultant. He authored numerous books that include *Introduction to Statistics via Resampling Methods and R/S-PLUS* and *Common Errors in Statistics (and How to Avoid Them)*. Email: drgood@statcourse.com.

powerful than the analysis of variance when samples are taken from non-normal distributions. For example, when the data in a four-sample, one-factor comparison are drawn from mixtures of normal distributions, 50% N(δ, 1) and 50% N(1+δ, 1), in an unbalanced design with 2, 3, 3, and 4 observations per cell, the permutation test was more powerful at the 10% level, a power of 86% against a shift in means of two units compared to 65% for the analysis of variance.

Unfortunately, the permutation test for interaction in a two-factor experimental design based on the set of all possible rearrangements among the cells is not exact. The residual errors are not exchangeable, nor are the p-values of such permutation tests for main effects and interactions independent of one another. Here is why:

Suppose the observations satisfy a linear model, $X_{ijm} = \mu + s_i + r_j + (sr)_{ij} + \varepsilon_{ijm}$ where the residual errors $\{\varepsilon_{ijm}\}$ are independent and identically distributed. To test the hypothesis of no interaction, first eliminate row and column effects by subtracting the row and column means from the original observations. That is, set

$$X'_{ijk} = X_{ijk} - \bar{X}_{i\_} - \bar{X}_{.j.} + \bar{X}_{\_\_},$$

where by adding the grand mean $\bar{X}_{\_}$, ensure the overall sum will be zero. Recall that

$$X'_{ijk} = X_{ijk} - \bar{X}_{i\_} - \bar{X}_{.j.} + \bar{X}_{\_}$$

or, in terms of the original linear model, that

$$X'_{ijk} = \varepsilon_{ijk} - \bar{\varepsilon}_{i\_} - \bar{\varepsilon}_{.j.} + \bar{\varepsilon}_{\_}.$$

However, this means that two residuals in the same row such as X'_{i11} and X'_{i23} will be correlated while residuals taken from different rows and columns will not be. Thus, the residuals are not exchangeable, a necessary requirement for tests based on a permutation distribution to be exact and independent of one another (see, for example, Good, 2002).

An alternative approach, first advanced by Salmaso and later published by Pesarin (2001) and Good (2002), is to restrict the permutation set to synchronized permutations in which, for example, an exchange between rows in one column is duplicated by exchanges between the same rows in all the other columns so as to preserve the exchangeability of the residuals.

The purpose of this article is to compare the power of ANOVA tests with those of permutation tests (synchronized and unsynchronized) when applied to two-factor experimental designs.

Methodology

Observations were drawn from one of the following three distributions:

1. Normal.
2. Weibull, because such distributions arise in reliability and survival analysis and cannot be readily transformed to normal distributions. A shape parameter of 1.5 was specified.
3. Contaminated normal, both because such mixtures of distributions are common in practice and because they cannot be readily transformed to normal distributions. In line with findings in an earlier article in this series, Good and Lunneborg (2006), we focused on the worst case distribution, a mixture of 70% N(0, 1) and 30% N(2, 2) observations.

Designs with the following effects were studied:

a)

+δ 0

+δ 0

b)

+δ 0

0 +δ

c)

+δ 0 … -δ

+δ 0 … - δ

d)

+δ 0 … 0 -δ

-δ 0 … 0 +δ

e)

0 +δ 0

0 +δ 0

0 +δ 0

f)

0 +δ 0

0 0 0

+δ 0 0

g)

+δ 0 0

0 +δ 0

0 0 +δ

h)

0 +δ 0 0 -δ

0 +δ 0 0 –δ

0 +δ 0 0 –δ

0 +δ 0 0 –δ

i)

1 δ 1 1 δ

1 δ 1 1 δ

1 δ 1 1 δ

1 δ 1 1 δ

To compare the results of the three methodologies, 1,000 data sets were generated at random for each design and each alternative (δ =0, 1, or 2). p-values for the permutation tests were obtained by Monte Carlo means using a minimum of 400 random (synchronized or unsynchronized) permutations per data set. The alpha level was set at 10%. (The exception being the 2x2 designs with 3 observations per cell where the highly discrete nature of the synchronized permutation distribution forced adoption of an 11.2% level.)

The simulations were programmed in R. Test results for the analysis of variance were derived using the anres() function. R code for the permutation tests and the data generators is posted at: http://statcourse.com/AnovPower.txt.

Results

Summary

In line with Jager's (1980) theoretical results, the analysis of variance (ANOV) applied to RxC experimental designs was found to yield almost exact tests even when data are drawn from mixtures of normal populations or from a Weibull distribution. This result holds whether the design is balanced or unbalanced. Of course, because the ANOVA tests for main effects and interaction share a common denominator - the within sum of squares - the resultant p-values are positively correlated. Thus a real non-zero main effect may be obscured by the presence of a spuriously significant interaction.

Although tests based on synchronized permutations are both exact and independent of one another, there are so few synchronized permutations with small samples that these tests lack power. For example, in a 2x2 design with 3 observations per cell, there are only 9 distinct values of each of the test statistics.

Fortunately, tests based on the entire set of permutations, unsynchronized as well as synchronized, prove to be almost exact. Moreover, these permutation tests for main effects and interaction are negatively correlated. The result is an increase in power if only one effect is present, but a loss in power if there are multiple effects. These permutation tests are more powerful than ANOVA tests when the data are drawn from mixtures of normal populations or from a Weibull distribution. They are as powerful, even with data drawn from normal distributions, with samples of $n \geq 5$ per cell.

2xK Design

In a 2x2 design with 3 observations per cell, restricting the permutation distribution to synchronized permutations means there are only 9 distinct values of each of the test statistics. The resultant tests lack power as do the tests based on synchronized permutations for 2x5 designs with as many as five observations per cell. For example, in a 2x4 design with four observations per cell, the synchronized permutation test had a power of 53% against a shift of two units when the data were drawn from a contaminated normal, while the power of the equivalent ANOVA test was 61%. As a result of these

negative findings, synchronized permutation tests were eliminated from further consideration.

In a balanced 2x2 design with 5 observations per cell, the powers of the ANOVA test and the traditional permutation test against a normal are equivalent. Against a contaminated normal or Weibull alternative, the permutation test is fractionally better. With only 3 observations per cell and a Weibull alternative with a doubling of scale, the permutation test is again fractionally superior.

In an unbalanced 2x2 design with 5 observations in each cell of the first column, and 3 observations in each cell of the second column, against a normal with a column effect of one unit (design a), ANOVA is markedly inferior with a power of 60% versus a power of 70% for the permutation test. Against a Weibull alternative with a doubling of the scale factor, the power of the ANOVA is 56%, while that of the permutation test is 71%. Noteworthy in this latter instance is that although there is no interaction term in design a, spurious interaction was recorded 18% of the time by the analysis of variance and 13% by permutation methods.

In a 2x5 design of form c with 3 observations per cell, the permutation test is several percentage points more powerful than ANOVA against both normal and contaminated normal alternatives.

3x3 Designs

When row, column, and interactions are all present as in design f, ANOVA is more powerful than the permutation test by several percentage points for all effects against both normal and contaminated normal alternatives. (See Table 1a, b.)

Table1a: Normal Alternative $\delta = 1$, 3 Observations Per Cell, Design f

Row-Column Interaction	
ANOVA	Permutation
187	139
178	138
344	316

Table 1b: Contaminated Normal Alternative $\delta = 2$, 3 Observations Per Cell, Design f

Row-Column Interaction	
ANOVA	Permutation
150	114
169	137
336	318

However, when a pure column effect (design e) or a pure interaction (design g) exists, the permutation test is superior to the analysis of variance by several percentage points. See, for example, Table 2.

Table 2: Contaminated Normal Alternative $\delta = 2$, 3 Observations Per Cell, Design g

Row-Column Interaction	
ANOVA	Permutation
115	70
108	70
461	529

4x5 Designs

The power against balanced designs of type h with four observations per cell of permutation and ANOVA tests are equivalent when the data is drawn from a normal distribution. The power of the permutation test is fractionally superior when the data is drawn from a mixed-normal distribution. Likewise, with a design of type i, the permutation test is several percentage points superior when the data is drawn from a Weibull distribution and the design is balanced. Synchronized permutations fared worst of all, their power being several percentage points below that provided by the analysis of variance.

When the design is unbalanced as in

4	4	4	4	4
4	4	4	4	4
2	3	4	5	3
2	3	4	5	3

the analysis of variance has the advantage in power over the permutation tests by several percentage points.

Discussion

Apart from 2xC designs, there appears to be little advantage to performing alternatives to the standard analysis of variance. The permutation tests are more powerful if only a single effect is present, but how often can this be guaranteed? Even with 2xC designs, the results reported here will be of little practical value until and unless permutation methods are incorporated in standard commercial packages. Wheeler suggests in a personal communication that if a package possesses a macro-language, a vector permutation command and an ANOVA routine, a permutation test for the multi-factor design can be readily assembled as follows:

1. Use the ANOVA command applied to the original data set to generate the sums of squares used in the denominators of the tests of the various effects.
2. Set up a loop and perform the following steps repeatedly:
    a. Rearrange the data.
    b. Use the ANOVA command applied to the rearranged data set to generate the sums of squares used in the denominators of the tests of the various effects.
    c. Compare these sums with the sums for the original data set.
3. Record the p-values as the percentage of rearrangements in which the new sum equaled or exceeded the value of the original.

References

David, F. N., & Johnson, N. L. (1951). The effect of non-normality on the power function of the f-test in the analysis of variance. *Biometrika*, 38, 43-57.

Good, P. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, *1*, 243-247.

Good, P. (2005). *Permutation, parametric, and bootstrap tests of hypotheses* (*3rd Ed.*). NY: Springer.

Good, P., & Lunneborg, C. E. (2005). Limitations of the analysis of variance. I: The one-way design. *Journal of Modern Applied Statistical Methods*, *5*(*1*), 41-43.

Jagers, P. (1980). Invariance in the linear model—an argument for chi-square and f in nonnormal situations. *Mathematische Operationsforschung und Statistik*, *11*, 455-464.

Pesarin, F. (2001). *Multivariate permutation tests*. NY: Wiley.

Salmaso, L. (2003). Synchronized permutation tests in 2^k factorial designs. *Communications in Statistics - Theory and Methods*, *32*, 1419-1438.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 360-376

1538 – 9472/09/$95.00

REGULAR ARTICLES

Examples of Computing Power for Zero-Inflated and Overdispersed Count Data

Suzanne R. Doyle
University of Washington

Examples of zero-inflated Poisson and negative binomial regression models were used to demonstrate conditional power estimation, utilizing the method of an expanded data set derived from probability weights based on assumed regression parameter values. SAS code is provided to calculate power for models with a binary or continuous covariate associated with zero-inflation.

Key words: Conditional power, Wald statistic, zero-inflation, over-dispersion, Poisson, negative binomial.

Introduction

Lyles, Lin and Williamson (2007) presented a simple method for estimating conditional power (i.e., power given a pre-specified covariate design matrix) for nominal, count or ordinal outcomes based on a given sample size. Their method requires fitting a regression model to an expanded data set using weights that represent response probabilities, given assumed values of covariate regression parameters. It has the flexibility to handle multiple binary or continuous covariates, requires only standard software and does not involve complex mathematical calculations. To estimate power, the variance-covariance matrix of the fitted model is used to derive a non-central chi square approximation to the distribution of the Wald statistic. This method can also be used to approximate power for the likelihood ratio test.

Lyles, et al. (2007) illustrated the method for a variety of outcome types and covariate patterns, and generated simulated data to demonstrate its accuracy. In addition to the proportional odds model and logistic regression, they included standard Poisson regression with one continuous covariate and negative binomial regression with one binary covariate. Both the Poisson and negative binomial regression models provide a common framework for the analysis of non-negative count data. If the model mean and variance values are the same (equi-dispersion), the one-parameter Poisson distribution can be appropriately used to model such count data. However, when the sample variance exceeds the sample mean (over-dispersion), the negative binomial distribution provides an alternative by using a second parameter for adjusting the variance independently of the mean.

Over-dispersion of count data can also occur when there is an excess proportion of zeros relative to what would be expected with the standard Poisson distribution. In this case, generalizations of the Poisson model, known as zero-inflated Poisson (ZIP) and ZIP(τ) (Lambert, 1992), are more appropriate when there is an excess proportion of zeros and equi-dispersion of the non-zero count data is present. These models provide a mixture of regression models: a logistic portion that accounts for the probability of a count of zero and a Poisson portion contributing to the frequency of positive counts. The ZIP model permits different covariates and coefficient values between the logistic and Poisson portions of the model. Alternatively, the ZIP(τ) model is suitable when covariates are the same and the logistic parameters are functionally related to the Poisson parameters.

Suzanne R. Doyle is a Biostatistician in the Alcohol and Drug Abuse Institute. Email: srdoyle@u.washington.edu.

With the ZIP and ZIP(τ) models, the non-zero counts are assumed to demonstrate equi-dispersion. However, if there is zero-inflation and non-zero counts are over-dispersed in relation to the Poisson distribution, parameter estimates will be biased and an alternative distribution, such as the zero-inflated negative binomial regression models, ZINB or ZINB(τ), are more appropriate (Greene, 1994). Similar to the zero-inflated Poisson models, ZINB allows for different covariates and ZINB(τ) permits the same covariates between the logistic portion for zero counts and the negative binomial distribution for non-zero counts.

In this study, the use of an expanded data set and the method of calculating conditional power as presented by Lyles, et al. (2007) is extended to include the ZIP, ZIP(τ), ZINB and ZINB(τ) models. Examples allow for the use of a binary or a normally-distributed continuous covariate associated with the zero-inflation. Simulations were conducted to assess the accuracy of calculated power estimates and example SAS software programs (SAS Institute, 2004) are provided.

Methodology

Model and Hypothesis Testing

Following directly from Lyles, et al. (2007), the response variable Y for non-continuous count data has J possible values (y_1, y_2, … , y_J), a design matrix $\mathbf{X}$, and a regression model in the form of

$$\log(\lambda_i) = \boldsymbol{\beta}'\mathbf{x}_i \qquad (1)$$

with an assumed Poisson distribution or negative binomial distribution, where i indexes independent subjects (i = 1, ... , N), $\mathbf{x}_i$ is a (1 x q) vector of covariates, and $\boldsymbol{\beta}$ is a (1 x q) vector of regression coefficients. Under the Poisson or negative binomial regression model, the probabilities can be specified for j = 1, … , J by

$$w_{ij} = \Pr(Y_i = y_j \mid \mathbf{X}_i = \mathbf{x}_i),\ y_j = 0, 1, \ldots, \infty \qquad (2)$$

Interest is in testing the hypothesis $H_0: \mathbf{H}\boldsymbol{\beta} = \mathbf{h}_0$ *versus* $H_A: \mathbf{H}\boldsymbol{\beta} \neq \mathbf{h}_0$, where $\mathbf{H}$ is an (h x q) matrix of full row rank and $\mathbf{h}_0$ an (h x 1) constant vector. The Wald test statistic is

$$W = (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}_0)'[\mathbf{H}\,\hat{var}(\hat{\boldsymbol{\beta}})\mathbf{H}']^{-1}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}_0) \qquad (3)$$

where $\hat{\boldsymbol{\beta}}$ contains unrestricted maximum likelihood estimates of $\boldsymbol{\beta}$. Under H_0, (3) is asymptotically distributed as central chi square with h degrees of freedom (χ^2_h).

For power calculations, under H_A, the Wald test statistic is asymptotically distributed as non-central $\chi^2_{h,(\eta)}$, where the non-centrality parameter η is defined as

$$\eta = (\mathbf{H}\boldsymbol{\beta} - \mathbf{h}_0)'[\mathbf{H}\,\hat{var}(\hat{\boldsymbol{\beta}})\mathbf{H}']^{-1}(\mathbf{H}\boldsymbol{\beta} - \mathbf{h}_0). \qquad (4)$$

Creating an Expanded Data Set and Computing Conditional Power

To estimate the conditional power given assumed values of N, $\mathbf{X}$ and $\boldsymbol{\beta}$, an expanded data set is first created by selecting a value of J for the number of possible values of Y with non-negligible probability for any specific $\boldsymbol{x}_i$, such that

$$\sum_{j=1}^{J} w_{ij} = \sum_{j=1}^{J} \Pr(Y_i = y_i \mid \mathbf{X}_i = \mathbf{x}_i) \approx 1 \qquad (5)$$

for all i. The sum in (5) should be checked for each unique value of $\boldsymbol{x}_i$. A reasonable threshold for the sum (e.g., > 0.9999) is suggested for sufficient accuracy (Lyles et al., 2007). Second, for each value of i = 1, ... , N, a data matrix with J rows is created with the weights w_{ij} in (2) being computed with the assumed values of $\boldsymbol{\beta}$. This data matrix with J rows is stacked N times vertically from i = 1, ... , N to form an expanded data set with NJ records. The resulting expanded data set can be based on the same number of J records for each value of i. However, J can vary with i, as long as the condition in (5) is satisfied.

When the expanded data set is correctly created maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ from maximizing the weighted log-likelihood should equal the assumed value of $\boldsymbol{\beta}$, and the matrix $\hat{var}(\hat{\boldsymbol{\beta}})$ will accurately reflect variability under the specific model allowing for power

calculations based on the Wald test of (3). For more detailed information and examples concerning model and hypothesis testing and creating an expanded data set with this method, see Lyles, et al. (2007).

Subsequent to fitting the model to the expanded data set, the non-centrality parameter η in (4) is derived. Power is then calculated as

$$\Pr(x^2_{h(\eta)} \geq \chi^2_{h,1-\alpha}) \qquad (6)$$

where $x^2_{h,1-\alpha}$ denotes the 100(1 - α) percentile of the central χ^2 distribution with *h* degrees of freedom. For testing a single regression coefficient, $\eta = \beta_k^2 / \hat{\sigma}_k^2$, where $\hat{\sigma}_k$ is the associated estimated standard error, with *h* = 1.

Zero-Inflated Poisson and Negative Binomial Models

Following Lambert (1992) for the zero-inflated Poisson (ZIP) regression model and Greene (1994) for the zero-inflated negative binomial (ZINB) regression model, the response Y_i is given by

$$Y_i \sim 0 \text{ with probability } \pi_i,$$

$$Y_i \sim \text{Poisson}(\lambda_i) \text{ for the ZIP model}$$

or

$$Y_i \sim \text{NegBin}(\lambda_i) \text{ for the ZINB model,}$$

with probability $1-\pi_i$, i = 1, ... , n for both models. For these models, the probability of zero counts is given by

$$\Pr(Y_i = 0) = \pi_i + (1-\pi_i)e^{-\lambda_i}. \qquad (7)$$

The probability of non-zero counts for the ZIP model is

$$\Pr(Y_i = y_j \mid \mathbf{X}_i = \mathbf{x}_i) = (1-\pi_i)\frac{e^{-\lambda_i}\lambda_i^{y_j}}{y_j!}, \quad (8)$$

and for the ZINB model is

$$\Pr(Y_i = y_j \mid \mathbf{X}_i = \mathbf{x}_i) =$$

$$(1-\pi_i)\frac{\Gamma(\kappa^{-1}+y_j)}{\Gamma(\kappa^{-1})y_j!}\left(\frac{\kappa\lambda_i}{1+\kappa\lambda_i}\right)^{y_j}\left(\frac{1}{1+\kappa\lambda_i}\right)^{1/\kappa}. \quad (9)$$

for y_j = 1, ... , ∞, where Γ is the gamma function. In contrast to the Poisson model with only one parameter, the negative binomial model has two parameters: λ (the mean, or shape parameter) and a scale parameter, κ, both of which are non-negative for zero-inflated models, and not necessarily an integer. Both π_i of the logistic model and λ_i of the Poisson model or negative binomial model depend on covariates through canonical link of the generalized linear model

$$\text{logit}(\pi_i) = \gamma' z_i$$

and

$$\log(\lambda_i) = \beta' x_i \qquad (10)$$

with $\gamma = (\gamma_0, \gamma_1, ..., \gamma_r)$ and $\beta = (\beta_0, \beta_1, ..., \beta_p)$. Because the covariates that influence π_i and λ_i are not necessarily the same, two different sets of covariate vectors, $z_i = (1, z_{i1}, ..., z_{ir})$ and $x_i = (1, x_{i1}, ..., x_{ip})$, are allowed in the model. Interpretation of the γ and β parameters is the same as the interpretation of the parameters from standard logistic and Poisson or negative binomial models, respectively.

If the same covariates influence π_i and λ_i, and if π_i can be written as a scalar multiple of λ_i, such that

$$\text{logit}(\pi_i) = -\tau\beta' x_i$$

and

$$\log(\lambda_i) = \beta' x_i \qquad (11)$$

then the ZIP and ZINB models described in (10) are called ZIP(τ) or ZINB(τ) models with an unknown scalar shape parameter τ (Lambert, 1992). When $\tau > 0$ zero inflation is less likely, and as $\tau \to 0$ zero inflation increases. Note that the number of parameters in the ZIP(τ) and ZINB(τ) models is reduced, providing a more parsimonious model than the ZIP and ZINB

models, and it may therefore be advantageous to use this model when appropriate.

With the ZIP and ZIP(τ) models, the weights for the expanded data set are calculated as

$$w_{ij} = \Pr(Y_i = y_j \mid \mathbf{X}_i = \mathbf{x}_i) = \pi_i I(y_i) + (1-\pi_i)\frac{e^{-\lambda_i}\lambda_i^{y_j}}{y_j!} \quad (12)$$

and for the ZINB and ZINB(τ) models, the weights for the expanded data set are

$$w_{ij} = \Pr(Y_i = y_j \mid \mathbf{X}_i = \mathbf{x}_i) = (\pi_i)I(y_i) + (1-\pi_i)\frac{\Gamma(\kappa^{-1}+y_j)}{\Gamma(\kappa^{-1})y_j!}\left(\frac{\kappa\lambda_i}{1+\kappa\lambda_i}\right)^{y_j}\left(\frac{1}{1+\kappa\lambda_i}\right)^{1/\kappa} \quad (13)$$

with y_j = 0, 1, ... , ∞, and where $I(y_i)$ is an indicator function taking a value of 1 if the observed response is zero $(y_i = 0)$ and a value of 0 if the observed response is positive $(y_i > 0)$.

Simulating the Negative Binomial Distribution

In simulating the negative binomial distribution, Lyles, et al. (2007) generated independent geometric random variates, under the constraint of only integer values for $1/\kappa$. In contrast, the negative binomial distribution in this study was simulated according to the framework provided by Lord (2006). This algorithm is based on the fact that the negative binomial distribution can be characterized as a Poisson-gamma mixture model (Cameron & Trivedi, 2006), it is consistent with the linear modeling approach used with this method of power calculation and also allows for non-integer values of $1/\kappa$. To calculate an outcome variable that is distributed as negative binomial, the following steps are taken:

1. Generate a mean value (λ_i) for observation *i* from a fixed sample population mean, $\lambda_i = \exp(\beta' x_i)$
2. Generate a value (ϕ_i) from a gamma distribution with the mean equal to 1 and the parameter $\delta = 1/\kappa$, $\phi_i = \Gamma(\delta, 1/\delta)$
3. Calculate the mean (θ_i) for observation *i*, $\theta_i = \lambda_i \times \phi_i$
4. Generate a discrete value (y_i) for observation *i* from a Poisson distribution with a mean θ_i, $y_i \sim \text{Poisson}(\theta_i)$.
5. Repeat steps 1 through 4 N times, where N is the number of observations or sample size.

Examples

Several examples are presented to illustrate the conditional power calculations of the ZIP, ZIP(τ), ZINB and ZINB(τ) models with a binary or continuous covariate related to the logistic portion accounting for the zero-inflation. Models were selected to demonstrate the effects of increased zero-inflation and over-dispersion on power estimates. Each model was fit by utilizing a weighted form of the general log-likelihood feature in SAS PROC NLMIXED (SAS Institute, 2004). Simulations under each model and the assumed joint covariate distributions were conducted to assess the accuracy of the power calculations. In situations where a reasonable solution could not be obtained with the generated data, the simulation data set was excluded from consideration and data generation was continued until 1,000 usable data sets were obtained for each model. A non-viable solution was generally due to non-convergence or extremely large standard errors.

In particular, the ZIP(τ) and ZINB(τ) models were the most problematic due to obtaining extremely large standard errors and parameter estimates of τ. In some situations it was obvious that a poor solution resulted, but in other instances it was not as clear that an unsatisfactory solution occurred. To avoid arbitrary decisions on which simulations to exclude, all data sets resulting in a value of τ outside of the boundaries of a 99% confidence interval (based on assumed regression parameter values) were deleted. A similar decision rule was used for the ZIP and ZINB models, eliminating data sets with values of γ_i, as

defined in (10) beyond their 99% confidence boundaries. The selection decision to discard data sets from consideration did not depend on the values of the regression parameter of interest to be statistically tested.

Simulation values presented are the average regression coefficient and the average standard error (calculated as the square root of the average error variance) out of the 1,000 generated data sets for the parameter estimates of each model. Simulation-based power was calculated as the proportion of Wald tests found statistically significant at $\alpha = .05$ out of 1,000 randomly generated data sets under each specific model considered. Appendices A through D provide SAS programming code to evaluate a large sample simulation for distributional characteristics, to construct an expanded data set and to calculate power for models with a binary covariate or a normally-distributed continuous covariate related to the zero-inflation.

To calculate the expanded data set, it was first necessary to choose the initial value of J for each value of x_i. This was done by generating a large simulated data set (N = 100,000 for each binary or continuous covariate in the model) based on the same parameter values of the model. To ensure that a reasonable threshold for the sum (e.g., > 0.9999) of the weights in (12) and (13) would be obtained, the initial value of J was increased in one unit integer increments until the maximum likelihood estimates for the parameters from maximizing the weighted log-likelihood equaled the assumed parameter values of the regression model. The large simulated data set also provided approximations to the population distributional characteristics of each model (mean, variance, and frequencies of each value of the outcome variable y_i) and estimates of the percents of zero-inflation.

ZIP(τ), ZIP, ZINB(τ) and ZINB Models with a Binary Variable for Zero Inflation

Model A-τ and Model B-τ, where τ = 2 and 1, are ZIP(τ) and ZIP models, respectively. Model A-τ is defined as

$$\text{logit}(\pi_i) = -\tau\beta_0 - \tau\beta_1 x \text{ and } \log(\lambda_i) = \beta_0 + \beta_1 x \quad (14)$$

where $\beta_0 = 0.6931$, $\beta_1 = -0.3567$, $\tau = 2$ and 1, and x is a binary variable with an equal number of cases coded 0 and 1. The regression coefficients were based on the rate ratio. That is, for the binary covariate x, from the rates of the two groups ($\lambda_1 = 2$ and $\lambda_2 = 1.4$), the regression coefficients are $\beta_0 = \log\lambda_1$ and $\beta_1 = \log(\lambda_2) - \log(\lambda_1)$. With this model, interest is in testing $H_0 : \beta_1 = 0$ *versus* $H_A : \beta_1 \neq 0$.

Model B-τ is defined as

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 z$$

and

$$\log(\lambda_i) = \beta_0 + \beta_1 z + \beta_2 x \quad (15)$$

where $\beta_0 = 0.6931$, $\beta_1 = -0.3567$, $\beta_2 = -0.3567$, $\gamma_0 = -\tau\beta_0$, $\gamma_1 = -\tau\beta_1$, $\tau = 2$ and 1, and x and z are binary covariates with an equal number of cases coded 0 and 1. In this particular example the regression coefficients for the logistic portion of the model (γ_0 and γ_1) are both a constant multiple of τ, although this is not a necessary requirement for the ZIP model. With this model, interest is in assessing $H_0 : \beta_2 = 0$ versus $H_A : \beta_2 \neq 0$.

The ZINB(τ) and ZINB models consisted of the same parameter estimates as the ZIP(τ) and ZIP models (Model A-τ and Model B-τ described above), but included two values of an extra scale parameter, $\kappa = 0.75$ and $\kappa = 1.50$. Sample sizes were based on obtaining conditional power estimates of approximately .95 for the regression coefficient tested, with τ = 2 for the ZIP and ZIP(τ) models, and for τ = 2 and $\kappa = 0.75$ for the ZINB and ZINB(τ) models. SAS code to evaluate a large sample simulation for distributional characteristics, to construct an expanded data set and to calculate power, for models with a binary covariate related to the zero-inflation are presented in Appendices A and B, for the Poisson and negative binomial regression models, respectively.

Results

ZIP(τ) Models

The results of the ZIP(τ) models presented at the top of Table 1 indicate that with a sample size of N = 212, when τ = 2, there is approximately .95 power and 27.0% zero-inflation for testing $H_A: \beta_1 \neq 0$. As τ decreases and therefore zero-inflation increases, the calculated power is reduced to 0.81 for approximately 37.5% estimated zero-inflation. In most cases, the simulated parameter and power estimates match the calculated values, except for a slight tendency for the simulated data to result in inflated average parameter estimates for the standard error σ_τ.

The outcomes for the ZIP models presented at the bottom of Table 1 show that with a sample size of N = 488, when τ = 2, there is approximately .95 power and 27.0% zero-inflation for testing $H_A: \beta_2 \neq 0$. Again, as τ decreases, the calculated power is reduced to approximately .90 and 37.5% estimated zero-inflation.

ZINB(τ) Models With A Binary Covariate Associated With The Zero-Inflation

The results of the ZINB(τ) models with a binary covariate associated with the zero-inflation, presented at the top of Table 2, indicate that with a sample size of N = 464, when $\kappa = 0.75$ and τ = 2, there is approximately .95 power and 27.0% zero-inflation for testing $H_A: \beta_1 \neq 0$. As τ decreases to 1, the calculated power is reduced to approximately .80 and 37.5% estimated zero-inflation. When over-dispersion of the non-zero counts increases ($\kappa = 1.50$), power is reduced to approximately .80 when τ = 2, and .59 when τ = 1.

In most cases, the simulated (Sim.) values and power estimates closely match the calculated (Cal.) parameters, except for a slight tendency for the simulated data to result in an inflated average standard error (σ_τ) associated with parameter estimates for τ, and slightly lower than expected values for the scale or over-dispersion parameter κ.

The results of the ZINB models presented at the bottom of Table 2 indicate that with a sample size of N = 928, when $\kappa = 0.75$ and τ = 2, there is approximately .95 power and 27.0% zero-inflation for testing $H_A: \beta_2 \neq 0$. Again, as τ decreases (τ = 1), the calculated power is reduced to approximately .90 with 37.5% estimated zero-inflation. Also, when over-dispersion of the non-zero counts increases ($\kappa = 1.50$), power is reduced to approximately .85 when τ = 2, and .77 when τ = 1. There is also the slight tendency of the simulated data to result in average inflated standard errors (σ_{γ_0} and σ_{γ_1}) for the parameter estimates of the logistic portion of the model involving zero-inflation (γ_0 and γ_1), and in decreased values for the scale or over-dispersion parameter κ than would be expected.

ZIP(τ), ZIP, ZINB(τ) and ZINB Models with a Continuous Variable for Zero-Inflation

Model C-τ and Model D-τ, where τ = 2 and 1, are ZIP(τ) and ZIP models, respectively. Model C-τ is defined as

$$\text{logit}(\pi_i) = -\tau\beta_0 - \tau\beta_1 z \text{ and } \log(\lambda_i) = \beta_0 + \beta_1 z \quad (16)$$

where β_0 = 0.5000, β_1 = -0.1500, τ = 2 and 1, and z is a continuous variable distributed as N(0,1). These are the same parameter estimates of β_0 and β_1 used by Lyles, et al. (2007) with their example of standard Poisson regression with one continuous covariate. With this ZIP(τ) model, interest is in assessing $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$. Model D-τ is defined as

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 z \text{ and } \log(\lambda_i) = \beta_0 + \beta_1 z + \beta_2 x \quad (17)$$

where β_0 = 0.5000, β_1 = -0.1500, β_2 = -0.3000, $\gamma_0 = -\tau\beta_0$, $\gamma_1 = -\tau\beta_1$, τ = 2 and 1, *x* is a binary variable with an equal number of cases coded 0 and 1, and z is a continuous variable distributed

Table 1: Parameter Estimates with a Binary Covariate for Zero-Inflation and Poisson Regression

	ZIP(τ) Models (N = 212)			
	Model A-2		Model A-1	
	Calculated	Simulated	Calculated	Simulated
τ	2.0000	2.0622	1.0000	1.0390
σ_τ	0.6169	0.7034	0.4286	0.4792
β_0	0.6931	0.6962	0.6931	0.6944
σ_{β_0}	0.0891	0.0894	0.0990	0.0995
β_1	-0.3567	-0.3565	-0.3567	-0.3535
σ_{β_1}	0.0989	0.1005	0.1256	0.1284
β_1 Power	.9502	.9450	.8106	.8080
	Estimated Zero-Inflation			
$x = 0$	20.24%		33.70%	
$x = 1$	33.79%		41.59%	
Total	27.02%		37.65%	

	ZIP Models (N = 488)			
	Model B-2		Model B-1	
	Calculated	Simulated	Calculated	Simulated
γ_0	-1.3863	-1.4075	-0.6931	-0.7109
σ_{γ_0}	0.2670	0.2828	0.1966	0.2020
γ_1	0.7134	0.7140	0.3567	0.3462
σ_{γ_1}	0.3707	0.3948	0.3023	0.3179
β_0	0.6931	0.6923	0.6931	0.6893
σ_{β_0}	0.0789	0.0793	0.0865	0.0871
β_1	-0.3567	-0.3551	-0.3567	-0.3654
σ_{β_1}	0.1237	0.1246	0.1331	0.1348
β_2	-0.3567	-0.3608	-0.3567	-0.3554
σ_{β_2}	0.0991	0.0995	0.1105	0.1110
β_2 Power	.9494	.9540	.8976	.8980
	Estimated Zero-Inflation			
$z = 0$	20.08%		33.41%	
$z = 1$	33.71%		41.58%	
Total	26.90%		37.49%	

Table 2: Parameter Estimates with a Binary Covariate for Zero-Inflation and Negative Binomial Regression

	ZINB(τ) Models (N = 464)							
	Model A-2				Model A-1			
	$\kappa = 0.75$		$\kappa = 1.50$		$\kappa = 0.75$		$\kappa = 1.50$	
	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.
τ	2.0000	2.0357	2.0000	1.9844	1.0000	0.9932	1.0000	1.0301
σ_τ	1.1833	1.7476	1.8512	2.5505	0.7365	0.9743	1.1022	1.8936
β_0	0.6931	0.6997	0.6931	0.7429	0.6932	0.7081	0.6932	0.7246
σ_{β_0}	0.1394	0.1390	0.2089	0.1944	0.1507	0.1525	0.2192	0.2184
β_1	-0.3567	-0.3546	-0.3567	-0.3696	-0.3567	-0.3624	-0.3567	-0.3755
σ_{β_1}	0.0991	0.1015	0.1282	0.1334	0.1266	0.1319	0.1636	0.1690
κ	0.7500	0.7257	1.5000	1.3386	0.7500	0.7203	1.5000	1.4175
σ_κ	0.2545	0.2625	0.5440	0.5055	0.2635	0.2792	0.5405	0.5746
β_1 Power	.9494	.9520	.7946	.8182	.8044	.8120	.5872	.6030
Estimated Zero-Inflation								
$x = 0$	20.24%		20.14%		33.42%		33.43%	
$x = 1$	33.48%		34.06%		41.54%		41.90%	
Total	26.86%		27.10%		37.48%		37.66%	

	ZINB Models (N = 928)							
	Model A-2				Model A-1			
	$\kappa = 0.75$		$\kappa = 1.50$		$\kappa = 0.75$		$\kappa = 1.50$	
	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.
γ_0	-1.3864	-1.4394	-1.3858	-1.3223	-0.6931	-0.7373	-0.6930	-0.6546
σ_{γ_0}	0.4858	0.9712	0.7643	1.1975	0.3241	0.4062	0.5034	0.6438
γ_1	0.7134	0.7361	0.7128	0.6702	0.3568	0.3714	-0.3566	0.3405
σ_{γ_1}	0.3703	0.7984	0.4691	0.8262	0.2702	0.3187	0.3102	0.3877
β_0	0.6931	0.6940	0.6932	0.7235	0.6932	0.6960	0.6932	0.7210
σ_{β_0}	0.1199	0.1212	0.1773	0.1745	0.1307	0.1328	0.1931	0.1919
β_1	-0.3567	-0.3566	-0.3567	-0.3678	-0.3566	-0.3635	-0.3567	-0.3624
σ_{β_1}	0.1250	0.1249	0.1501	0.1482	0.1347	0.1344	0.1618	0.1596
β_2	-0.3567	-0.3584	-0.3567	-0.3545	-0.3567	-0.3584	-0.3567	-0.3538
σ_{β_2}	0.0992	0.0991	0.1198	0.1193	0.1096	0.1092	0.1318	0.1308
κ	0.7500	0.7380	1.4999	1.4080	0.7500	0.7417	1.4999	1.4173
σ_κ	0.2216	0.2333	0.4897	0.4968	0.2508	0.2611	0.5321	0.5606
β_2 Power	.9491	.9379	.8455	.8460	.9023	.9050	.7723	.7730
Estimated Zero-Inflation								
$z = 0$	20.22%		20.14%		33.51%		33.36%	
$z = 1$	33.74%		33.82%		41.70%		41.73%	
Total	26.98%		27.00%		37.60%		37.54%	

Note: Cal. indicates calculated values, and Sim. indicates simulated values.

Table 3: Parameter Estimates with a Continuous Covariate for Zero-Inflation and Poisson Regression

	ZIP(τ) Models (N = 302)			
	Model C-2		Model C-1	
	Calculated	Simulated	Calculated	Simulated
τ	2.0000	2.0411	1.0000	1.0566
σ_τ	0.5460	0.6017	0.3848	0.4254
β_0	0.5000	0.5004	0.5000	0.4963
σ_{β_0}	0.0625	0.0628	0.0685	0.0688
β_1	-0.1500	-0.1501	-0.1500	-0.1519
σ_{β_1}	0.0416	0.0421	0.0525	0.0529
β_1 Power	.9501	.9500	.8152	.8120
	Estimated Zero-Inflation			
Total	27.39%		37.99%	
	ZIP Models (N = 694)			
	Model D-2		Model D-1	
	Calculated	Simulated	Calculated	Simulated
γ_0	-1.0000	-1.0154	-0.5000	-0.5060
σ_{γ_0}	0.1521	0.1581	0.1253	0.1283
γ_1	0.3000	0.3057	0.1500	0.1492
σ_{γ_1}	0.1513	0.1563	0.1241	0.1271
β_0	0.5000	0.4979	0.5000	0.4953
σ_{β_0}	0.0610	0.0613	0.0662	0.0667
β_1	-0.1500	-0.1517	-0.1500	-0.1496
σ_{β_1}	0.0493	0.0494	0.0532	0.0536
β_2	-0.3000	-0.2992	-0.3000	-0.2987
σ_{β_2}	0.0832	0.0834	0.0925	0.0930
β_2 Power	.9501	.9510	.9003	.9010
	Estimated Zero-Inflation			
Total	27.32%		37.83%	

Table 4: Parameter Estimates with a Continuous Covariate for Zero-Inflation and Negative Binomial Regression

	ZINB(τ) Models (N = 648)							
	Model C-2				Model C-1			
	κ = 0.75		κ = 1.50		κ = 0.75		κ = 1.50	
	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.
τ	2.0000	1.9793	2.0000	2.0477	1.0000	1.0284	1.0000	1.0499
σ_τ	1.0565	1.3623	1.6377	2.3039	0.6816	0.9086	1.0280	1.7596
β_0	0.5000	0.5179	0.5000	0.5217	0.5000	0.5111	0.5000	0.5185
σ_{β_0}	0.0991	0.0984	0.1475	0.1460	0.1040	0.1059	0.1509	0.1571
β_1	-0.1500	-0.1537	-0.1500	-0.1527	-0.1500	-0.1504	-0.1500	-0.1562
σ_{β_1}	0.0416	0.0429	0.0533	0.0556	0.0530	0.0541	0.0680	0.0712
κ	0.7500	0.7390	1.5000	1.4247	0.7500	0.7297	1.5000	1.4372
$\sigma\kappa$	0.2224	0.2335	0.4733	0.4763	0.2331	0.2436	0.4832	0.5209
β_1 Power	.9501	.9470	.8035	.8110	.8079	.8130	.5972	.5980
Estimated Zero-Inflation								
Total	27.27%		27.46%		37.67%		37.90%	
	ZINB Models (N = 1324)							
	Model D-2				Model D-1			
	κ = 0.75		κ = 1.50		κ = 0.75		κ = 1.50	
	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.	Cal.	Sim.
γ_0	-1.0000	-1.0242	-1.0001	-0.9989	-0.5000	-0.5025	-0.5001	-0.5026
σ_{γ_0}	0.3105	0.3867	0.4902	0.5990	0.2386	0.2635	0.3751	0.4525
γ_1	0.3000	0.3066	0.3001	0.3136	0.1500	0.1554	0.1500	0.1554
σ_{γ_1}	0.1466	0.1757	0.1794	0.2260	0.1118	0.1197	0.1282	0.1528
β_0	0.5000	0.5024	0.5000	0.5176	0.5000	0.5034	0.5000	0.5090
σ_{β_0}	0.0967	0.0964	0.1442	0.1402	0.1049	0.1062	0.1570	0.1581
β_1	-0.1500	-0.1516	-0.1500	-0.1485	-0.1500	-0.1497	-0.1500	-0.1524
σ_{β_1}	0.0512	0.0508	0.0620	0.0615	0.0555	0.0555	0.0672	0.0670
β_2	-0.3000	-0.3050	-0.3000	-0.3020	-0.3000	-0.3012	-0.3000	-0.2988
σ_{β_2}	0.0832	0.0831	0.1005	0.1001	0.0918	0.0917	0.1104	0.1101
κ	0.7500	0.7370	1.5000	1.4476	0.7500	0.7392	1.5001	1.4665
$\sigma\kappa$	0.1868	0.1900	0.4101	0.4042	0.2032	0.2129	0.4479	0.4739
β_2 Power	.9501	.9530	.8473	.8420	.9046	.8980	.7756	.7710
Estimated Zero-Inflation								
Total	27.32%		27.37%		37.82%		37.79%	

Note: *Cal*. indicates calculated values, and Sim. indicates simulated values.

as N(0,1). With this ZIP model, interest is in testing $H_0: \beta_2 = 0$ versus $H_A: \beta_2 \neq 0$.

The ZINB(τ) and ZINB models consisted of the same parameter estimates as the ZIP(τ) and ZIP models (Model C-τ and Model D-τ), but included an extra scale parameter, $\kappa = 0.75$ and $\kappa = 1.50$. SAS programming code to evaluate a large sample simulation for distributional characteristics, to construct an expanded data set, and to calculate power, for models with a continuous covariate related to the zero-inflation are presented in Appendices C and D, for the Poisson and negative binomial regression models, respectively.

The results of the ZIP(τ) and ZIP models with a continuous covariate for zero-inflation are presented in Table 3. As before, when τ decreases, based on the same sample size and value of the regression coefficient tested, the calculated power is reduced, and there is also a slight tendency for the simulated data to result in inflated average parameter estimates for the standard error (σ_τ) with the ZIP(τ) models, and with inflated average parameter estimates of the standard errors for the logistic portion involving zero-inflation (σ_{γ_0} and σ_{γ_1}) with the ZIP models.

The results of the ZINB(τ) and ZINB models with a continuous covariate for zero-inflation are presented in Table 4. Similar to the results previously presented, based on the same sample size and value of the regression coefficient tested, when τ decreases and/or when overdispersion of the non-zero counts increases, the calculated power is reduced There is a slight tendency for simulated data to result in inflated average standard errors (σ_τ) for the parameter estimates of τ with the ZINB(τ) models, and with inflated average standard errors (σ_{γ_0} and σ_{γ_1}) for the logistic portion involving zero-inflation (γ_0 and γ_1) with the ZINB models.

Conclusion

Examples of ZIP, ZIP(τ), ZINB and ZINB(τ) models were used to extend the method of estimating conditional power presented by Lyles, et al. (2007) to zero-inflated count data. Utilizing the variance-covariance matrix of the model fitted to an expanded data set, power was estimated for the Wald statistic. Although not presented here, this method can also be used to approximate power based on the likelihood ratio test. Overall, with the same sample size and parameter value of the estimate of interest to be tested with the Wald test statistic, results indicated a decrease in power as the percent of zero-inflation and/or over-dispersion increased. This trend was particularly more noticeable for the ZIP(τ) and ZINB(τ) models. Calculated power estimates indicate if the percent of zero-inflation or over-dispersion is underestimated, a loss of assumed power in the statistical test will result.

To estimate power for zero-inflated count data it is necessary to select a value of τ for the ZIP(τ) and ZINB(τ) models or values of the regression coefficients associated with the logistic portion in the ZIP and ZINB models (i.e., γ_0 and γ_1) to produce the correct assumed proportion of zero-inflation. But in practice, these parameter values may be unknown or difficult to estimate. Generating a large simulated data set iteratively until the expected percent of zero-inflation occurs can aid the researcher in obtaining approximations to the population distributional characteristics of model and estimation of the parameter values associated with zero-inflation can be improved.

References

Cameron, A. C., & Trivedi, P. K. (2006). *Regression analysis of count data*. New York: Cambridge University Press.

Greene, W. H. (1994). *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. Stern School of Business, New York University, Dept. of Economics Working Paper, No. EC-94-10.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1-14.

Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*, *38*, 751-766.

Lyles, R. H., Lin, H-M., & Williamson J. M. (2007). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine*, 26**,** 1632-1648.

SAS Institute, Inc. (2004). *SAS/STAT 9.1 User's Guide*. SAS Institute Inc: Cary, NC.

Appendix A:
SAS Code with a Binary Covariate for Zero-Inflation and Poisson Regression

Step 1: Evaluate a large sample simulation for distributional characteristics.

ZIP(τ)

```
data ziptau1; seed = 12345;
lambda1 = 2; lambda2 = 1.4; tau = 2;
n = 100000;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
do x = 0 to 1; do i = 1 to n;
lambda = exp(beta0 + beta1*x);
prob_0 = exp(-tau*beta0 - tau*beta1*x)/
(1 + exp(-tau*beta0 - tau*beta1*x));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed,lambda);
if zero_inflate = 0 then yPoisson = y;
else yPoisson = .;
output; end; end;
proc sort; by x;
proc freq; tables y zero_inflate; by x; run;
proc freq; tables zero_inflate; run;
proc means mean var n; var y yPoisson;
by x; run;
```

ZIP

```
data zip1; seed = 12345;
lambda1 = 2; lambda2 = 1.4; tau = 2;
n = 100000;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
beta2 = beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
do x = 0 to 1; do z = 0 to 1; do i = 1 to n;
lambda = exp(beta0 + beta1*z + beta2*x);
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y=0;
else y = ranpoi(seed,lambda);
if zero_inflate = 0 then yPoisson = y;
else yPoisson=.;
output; end; end; end;
proc sort; by x z;
proc freq; tables y zero_inflate; by x z; run;
proc means mean var n; var y yPoisson;
by x z; run;
proc sort; by z;
proc freq; tables zero_inflate; by z; run;
proc freq; tables zero_inflate; run;
```

Step 2: Construct an expanded data set to approximate conditional power.

ZIP(τ)

```
data ziptau2;
lambda1 = 2; lambda2 = 1.4; tau = 2;
totaln = 212; numgroups = 2;
n = totaln/numgroups;
increment = 10;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
do x = 0 to 1;
if x = 0 then j = 13; if x = 1 then j = 9;
do i = 1 to n;
lambda = exp(beta0 + beta1*x);
prob_0 = exp(-tau*beta0 - tau*beta1*x)/
(1 + exp(-tau*beta0 - tau*beta1*x));
do y = 0 to j + increment;
if y = 0 then w = prob_0 + (1-prob_0) *(exp(-
lambda)*lambda**y)/gamma(y + 1);
if y > 0 then w = (1-prob_0)*(exp
(-lambda)*lambda**y)/gamma(y + 1);
output; end; end; end;
proc nlmixed tech=dbldog cov;
parameters t=3 b0=0 b1=0;
p0 = exp(-t*b0 - t*b1*x)/(1 + exp(-t*b0 -
t*b1*x)); mu = exp(b0 + b1*x);
if y = 0 then do;
ll = (log(p0 + (1 - p0)*exp(-mu))); end;
if y > 0 then do;
ll = (log(1 - p0) + y*log(mu) - lgamma(y + 1) -
mu); end; loglike = w*ll;
model y ~ general(loglike); run;
```

ZIP

```
data zip2;
lambda1 = 2; lambda2 = 1.4; tau = 2;
totaln = 488; numgroups = 4;
```

```
n = totaln/numgroups;
increment = 10;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
beta2 = beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
do x = 0 to 1; do z =0 to 1;
if x = 0 and z = 0 then j = 13;
if x = 0 and z = 1 then j = 9;
if x = 1 and z = 0 then j = 8;
if x = 1 and z = 1 then j = 7;
do I = 1 to n;
lambda = exp(beta0 + beta1*z + beta2*x);
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
do y = 0 to j + increment ;
if y = 0 then w = prob_0 + (1-prob_0)*
(exp(-lambda)*lambda**y)/gamma(y + 1);
if y > 0 then w = (1-prob_0)*(exp(-lambda)
*lambda**y)/gamma(y + 1);
output; end; end; end; end;
proc nlmixed tech=dbldog cov;
parameters g0=0 g1=0 b0=0 b1=0 b2=0;
p0 = exp(g0 + g1*z)/(1 + exp(g0 + g1*z));
mu = exp(b0 + b1*z + b2*x);
if y = 0 then do;
ll = (log(p0 + (1 - p0)*exp(-mu))); end;
if y > 0 then do;
ll = (log(1 - p0) + y*log(mu) - lgamma(y + 1) -
mu); end; loglike = w*ll;
model y ~ general(loglike); run;
```

Step 3: Calculate power.

```
data power; estimate = -0.3567; standerr =
0.0989;
eta = (estimate**2)/(standerr**2); critvalue =
cinv(.95,1);
power = 1-probchi(critvalue,1,eta); proc print;
var eta power; run;
```

Appendix B:
SAS Programming Code with a Binary Covariate for Zero-Inflation and Negative Binomial Regression

Step 1: Evaluate a large sample simulation for distributional characteristics.

ZINB(τ)

```
data zinbtau1; seed = 12345;
lambda1 = 2; lambda2 = 1.4; tau = 2;
n = 100000;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
kappa = .75; delta = 1/kappa;
do x = 0 to 1; do i = 1 to n;
lambda = exp(beta0 + beta1*x);
phi = 1/delta*rangam(seed,delta);
theta = lambda*phi;
prob_0 = exp(-tau*beta0 - tau*beta1*x)/
(1 + exp(-tau*beta0 - tau*beta1*x));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed,theta);
if zero_inflate = 0 then yPoisson = y;
else yPoisson = .; output; end; end;
proc sort; by x;
proc freq; tables y zero_inflate; by x; run;
proc freq; tables zero_inflate; run;
proc means mean var max; var y yPoisson;
by x; run;
proc means mean var n; var y yPoisson; run;
```

ZINB

```
data zinb1; seed = 12345;
lambda1 = 2; lambda2 = 1.4; tau = 2;
n = 100000; kappa = .75; delta = 1/kappa;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
beta2 = beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
do x = 0 to 1; do z = 0 to 1; do i = 1 to n;
lambda = exp(beta0 + beta1*z + beta2*x);
phi = 1/delta*rangam(seed,delta);
theta = lambda*phi;
prob_0 =exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed, theta);
if zero_inflate = 0 then yPoisson = y;
else yPoisson = .; output; end; end; end;
proc sort; by x z;
proc freq; tables y zero_inflate; by x z; run;
proc means mean var max n;
var y yPoisson; by x z; run;
proc sort; by z;
proc freq; tables y zero_inflate; by z; run;
proc freq; tables y zero_inflate; run;
```

Step 2: Construct an expanded data set to approximate conditional power.

ZINB(τ)

```
data zinbtau2;
lambda1 = 2; lambda2 = 1.4; tau = 2;
totaln = 464; numgroups = 2; kappa = .75;
n = totaln/numgroups; increment = 8;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
do x = 0 to 1;
if x = 0 then j = 29; if x = 1 then j = 20;
do i = 1 to n;
lambda = exp(beta0 + beta1*x);
prob_0 = exp(-tau*beta0 - tau*beta1*x)/
(1 + exp(-tau*beta0 - tau*beta1*x));
do y = 0 to j + increment;
if y = 0 then w = prob_0 + (1-prob_0) *
gamma(kappa**-1 + y)/
(gamma(kappa**-1)*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
if y > 0 then w = (1-prob_0)* gamma(kappa**-1
+ y)/(gamma(kappa**-1)*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
output; end; end; end;
proc nlmixed tech=dbldog cov;
parameters t=3 b0=0 b1=0 k=1;
p0 = exp(-t*b0 - t*b1*x)/(1 + exp(-t*b0
- t*b1*x)); mu = exp(b0 + b1*x);
if y = 0 then do;
ll = p0 + (1-p0)*exp(-(y+(1/k))* log(1+k*mu));
end;
if y > 0 then do;
ll = (1-p0)*exp(lgamma(y+(1/k)) - lgamma(y+1)
- lgamma(1/k) + y*log(k*mu) - (y + (1/k)) *
log(1 + k*mu)); end;
loglike = w * log(ll);
model y ~ general(loglike); run;
```

ZINB

```
data zinb2;
lambda1 = 2; lambda2 = 1.4; tau = 2;
totaln = 928; numgroups=4; kappa = .75;
n = totaln/numgroups; increment = 5;
beta0 = log(lambda1);
beta1 = log(lambda2) - log(lambda1);
beta2 = beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
do x = 0 to 1; do z = 0 to 1;
if x = 0 and z = 0 then j = 29;
if x = 0 and z = 1 then j = 20;
if x = 1 and z = 0 then j = 21;
if x = 1 and z = 1 then j = 14;
do i = 1 to n;
lambda = exp(beta0 + beta1*z + beta2*x);
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
do y = 0 to j + increment;
if y = 0 then w = prob_0 + (1-prob_0) *
gamma(kappa**-1 + y)/(gamma(kappa**-
1)*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
if y > 0 then w = (1-prob_0)* gamma(kappa**-1
+ y)/(gamma(kappa**-1)*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
output; end; end; end; end;
proc nlmixed tech=dbldog cov;
parameters g0=0 g1=0 b0=0 b1=0 b2=0 k=1;
p0 = exp(g0 + g1*z) / (1 + exp(g0
+ g1*z)); mu = exp(b0 + b1*z + b2*x);
if y = 0 then do;
ll = p0 + (1-p0)*exp(-(y+(1/k))* log(1+k*mu));
end;
if y > 0 then do;
ll = (1-p0)*exp(lgamma(y+(1/k)) - lgamma(y+1)
- lgamma(1/k) + y*log(k*mu) - (y + (1/k)) *
log(1 + k*mu)); end;
loglike = w * log(ll);
model y ~ general(loglike); run;
```

Step 3: Calculate power.

```
data power; estimate = -0.3567; standerr =
0.0991;
eta = (estimate**2)/(standerr**2); critvalue =
cinv(.95,1);
power = 1 - probchi(critvalue,1,eta); proc print;
var eta power; run;
```

Appendix C:
SAS Code with a Continuous Covariate for Zero-Inflation and Poisson Regression

Step 1: Evaluate a large sample simulation for distributional characteristics.

ZIP(τ)
```
data ziptau3; seed = 12345;
tau = 2; n = 100000;
beta0 = .50; beta1 = -.15;
do i = 1 to n;
z = rannor(seed);
lambda = exp(beta0 + beta1*z);
prob_0 = exp(-tau*beta0 - tau*beta1*z)/
(1 + exp(-tau*beta0 - tau*beta1*z));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed, lambda);
if zero_inflate = 0 then yPoisson=y;
else yPoisson = .;
output; end;
proc freq; tables y zero_inflate; run;
proc means mean var n; var y yPoisson; run;
```

ZIP
```
data zip3; seed = 12345;
tau = 2; n = 100000;
beta0 = .50; beta1 = -.15;
beta2 = 2 * beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
do x = 0 to 1; do i = 1 to n;
z = rannor(seed);
lambda = exp(beta0 + beta1*z + beta2*x);
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed, lambda);
if zero_inflate = 0 then yPoisson=y;
else yPoisson = .;
output; end; end;
proc freq; tables y zero_inflate; run;
proc sort; by x;
proc freq; tables y; by x; run;
proc means mean var n; var y yPoisson;
by x; run;
```

Step 2: Construct an expanded data set to approximate conditional power.

ZIP(τ)
```
data ziptau4;
tau = 2; n = 302; j = 11;
beta0 = .50; beta1 = -.15;
increment = 10;
do i = 1 to n;
z = probit((i - 0.375)/( n + 0.25));
lambda = exp(beta0 + beta1*z);
prob_0 = exp(-tau*beta0 - tau*beta1*z)/
(1 + exp(-tau*beta0 - tau*beta1*z));
do y = 0 to j + increment;
if y = 0 then w = prob_0 + (1-prob_0) *(exp(-
lambda)*lambda**y)/gamma(y+1);
if y > 0 then w = (1-prob_0)*(exp
(-lambda)*lambda**y)/gamma(y+1);
output; end; end;
proc nlmixed tech=dbldog cov;
parameters t=3 b0=0 b1=0;
p0 = exp(-t*b0 - t*b1*z)/(1 + exp(-t*b0
- t*b1*z)); mu = exp(b0 + b1*z);
if y = 0 then do;
ll = (log(p0 + (1-p0)*exp(-mu))); end;
if y > 0 then do;
ll = (log(1-p0) + y*log(mu) - lgamma(y+1) -
mu); end; loglike = w * ll;
model y ~ general(loglike); run;
```

ZIP
```
data zip4;
tau = 2; totaln = 694; numgroups=2;
n = totaln/numgroups; increment = 10;
beta0 = .50; beta1 = -.15;
beta2 = 2* beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
do x = 0 to 1;
if x = 0 then j = 11; if x = 1 then j = 9;
do i = 1 to n;
z = probit((i -0.375)/(n + 0.25));
lambda = exp(beta0 + beta1*z + beta2*x);
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
do y = 0 to j + increment ;
if y = 0 then w = prob_0 + (1-prob_0) *(exp(-
lambda)*lambda**y)/gamma(y+1);
if y > 0 then w = (1-prob_0)*(exp
(-lambda)*lambda**y)/gamma(y+1);
output; end; end; end;
proc nlmixed tech=dbldog cov;
parameters g0=0 g1=0 b0=0 b1=0 b2=0;
p0 = exp(g0 + g1*z)/(1 + exp(g0 + g1*z));
mu = exp(b0 + b1*z + b2*x);
if y = 0 then do;
ll = (log(p0 + (1-p0)*exp(-mu))); end;
if y > 0 then do;
ll = (log(1-p0) + y*log(mu) - lgamma(y+1)-
mu); end; loglike = w * ll;
```

```
model y ~ general(loglike); run;
```
Step 3: Calculate power.

```
data power; estimate = -0.1500; standerr =
0.0416;
eta = (estimate**2)/(standerr**2);
critvalue=cinv(.95,1);
power=1-probchi(critvalue,1,eta); proc print; var
eta power; run;
```

Appendix D:
SAS Programming Code with a Continuous Covariate for Zero-Inflation and Negative Binomial Regression

Step 1: Evaluate a large sample simulation for distributional characteristics.

ZINB(τ)

```
data zinbtau3; seed = 12345;
tau = 2; n = 100000;
beta0 = .50; beta1 = -.15;
kappa = .75; delta = 1/kappa;
do i = 1 to n;
z = rannor(seed);
lambda = exp(beta0 + beta1*z);
phi = 1/delta*rangam(seed,delta);
theta = lambda*phi;
prob_0 = exp(-tau*beta0 - tau*beta1*z)/
(1 + exp(-tau*beta0 - tau*beta1*z));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed,theta);
if zero_inflate=0 then yPoisson=y;
else yPoisson=.; output; end;
proc freq; tables y zero_inflate; run;
proc means mean var max n; var y yPoisson;
run;
```

ZINB

```
data zinb3; seed = 12345;
tau = 2; n = 100000;
beta0 = .50; beta1 = -.15;
beta2 = 2 * beta1;
gamma0 = -tau*beta0;
gamma1 = -tau*beta1;
kappa = .75; delta = 1/kappa;
do x = 0 to 1;
do i = 1 to n;
z = rannor(seed);
lambda = exp(beta0 + beta1*z + beta2*x);
phi = 1/delta*rangam(seed,delta);
theta = lambda*phi;
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
zero_inflate = ranbin(seed,1,prob_0);
if zero_inflate = 1 then y = 0;
else y = ranpoi(seed,theta);
if zero_inflate=0 then yPoisson=y;
else yPoisson=.; output; end; end;
proc sort; by x;
proc freq; tables y zero_inflate; by x; run;
proc freq; tables y zero_inflate; run;
proc means mean var max n; var y yPoisson; by
x; run;
proc means mean var n; var y yPoisson; run;
```

Step 2: Construct an expanded data set to approximate conditional power.

ZINB(τ)

```
data zinbtau4;
tau = 2; n = 648;
beta0 = .5; beta1 = -.15;
kappa = .75; j = 23; increment = 7;
do i = 1 to n;
z = probit((i - 0.375)/( n + 0.25));
lambda = exp(beta0 + beta1*z);
prob_0 = exp(-tau*beta0 - tau*beta1*z)/
(1 + exp(-tau*beta0 - tau*beta1*z));
do y = 0 to j + increment;
if y = 0 then w = prob_0 + (1-prob_0) *
gamma(kappa**-1 + y)/(gamma(kappa**-
1)*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
if y > 0 then w = (1-prob_0)*
gamma(kappa**-1 + y)/(gamma(kappa**-1)
*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
output; end; end;
proc nlmixed tech=dbldog cov;
parameters t=3 b0=0 b1=0 k=1;
p0 = exp(-t*b0 - t*b1*z)/(1 + exp(-t*b0
- t*b1*z)); mu = exp(b0 + b1*z);
if y = 0 then do;
ll = p0 + (1-p0)*exp(-(y+(1/k))* log(1+k*mu));
end;
if y > 0 then do;
ll = (1-p0)*exp(lgamma(y+(1/k)) - lgamma(y+1)
- lgamma(1/k) + y*log(k*mu)- (y + (1/k)) *
log(1 + k*mu)); end;
```

```
loglike = w * log(ll);
model y ~ general(loglike); run;

ZINB
data zinb4;
totaln = 1324; numgroups = 2;
n = totaln/numgroups; tau = 2;
beta0 = .5; beta1 = -.15; beta2 = 2*beta1;
gamma0 = -tau*beta0; gamma1 = -tau*beta1;
kappa = .75; increment = 5;
do x = 0 to 1;
if x = 0 then j = 23; if x = 1 then j = 19;
do i = 1 to n;
z = probit((i - 0.375)/( n + 0.25));
lambda = exp(beta0 + beta1*z + beta2*x);
prob_0 = exp(gamma0 + gamma1*z)/
(1 + exp(gamma0 + gamma1*z));
do y = 0 to j + increment;
if y = 0 then w = prob_0 + (1-prob_0) *
gamma(kappa**-1 + y)/
(gamma(kappa**-1) *gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
if y > 0 then w = (1-prob_0)* gamma(kappa**-1
+ y)/(gamma(kappa**-1)*gamma(y+1))*
((kappa*lambda/(1 + kappa*lambda))**y)
*((1/(1 + kappa*lambda))**(1/kappa));
output; end; end; end;
proc nlmixed tech=dbldog cov;
parameters g0=0 g1=0 b0=0 b1=0 b2=0;
p0 = exp(g0 + g1*z)/(1 + exp(g0 + g1*z));
mu = exp(b0 + b1*z + b2*x);
if y = 0 then do;
ll = p0 + (1-p0)*exp(-(y+(1/k))* log(1+k*mu));
end;
if y > 0 then do;
ll = (1-p0)*exp(lgamma(y+(1/k)) - lgamma(y+1)
- lgamma(1/k) + y*log(k*mu)- (y + (1/k)) *
log(1 + k*mu)); end;
loglike = w * log(ll);
model y ~ general(loglike); run;
```

Step 3: Calculate power.

```
data power; estimate = -0.1500; standerr =
0.0416;
eta = (estimate**2)/(standerr**2);
critvalue=cinv(.95,1);
power=1-probchi(critvalue,1,eta); proc print; var
eta power; run;
```

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 377-383

1538 – 9472/09/$95.00

Assessing Trends: Monte Carlo Trials with Four Different Regression Methods

Daniel R. Thompson
Florida Department of Health

Ordinary Least Squares (OLS), Poisson, Negative Binomial, and Quasi-Poisson Regression methods were assessed for testing the statistical significance of a trend by performing 10,000 simulations. The Poisson method should be used when data follow a Poisson distribution. The other methods should be used when data follow a normal distribution.

Key words: Monte Carlo, simulation, Ordinary least squares regression, Poisson regression, negative binomial regression, Quasi-Poisson regression.

Introduction

In the analysis of trend data, the key question is whether the trend reflects a true change or, alternatively, random variation. Statistical methods can be used to assess the probability that a trend has occurred due to chance. One approach is to use regression techniques to calculate the slope of the line that best fits the trend. If the slope of the line is significantly different from the flat line slope of zero, the trend is assumed to be non-random.

Disease and mortality rates generally change exponentially over time and are therefore linear in terms of the natural logarithm of the rate. Consequently, methods based on the slope of a straight line can be used to examine the natural logarithm of rates over time. The slope of the line that best fits the trend of the logarithm of the rates can also be used to calculate the estimated annual percent change (EAPC). This is explained in more detail on the National Cancer Institute internet web page under the Surveillance, Epidemiology and End Results program (SEER) (http://seer.cancer.gov/seerstat/WebHelp/Trend_Algorithms.htm).

Several commonly used methods for assessing the statistical significance of trends exist. These methods differ in the assumptions made about the distribution of the data and in the way the slope is calculated. The Poisson regression method assumes the numerator and denominator data for the rates follow a Poisson distribution and the variances are assumed to be equal to the means. The dependent variable is the natural logarithm of the numerators with the natural logarithm of the denominators used as an offset (Dunteman & Ho, 2006). This method has been used by Liu et al to analyze trends in stroke deaths in Japan (Liu, Ikeda & Yamori, 2006); by Botha et al to analyze trends in breast cancer deaths in Europe (Botha, et al., 2001) and by Lieb et al to analyze HIV/AIDS diagnosis trends in Florida (Lieb, et al., 2007).

The Quasi-Poisson and Negative Binomial regression methods are similar to the Poisson regression method but these methods do not assume the variances are equal to the means. For more information on the Quasi-Poisson and Negative Binomial methods see Wolfram Mathworld (http://mathworld.wolfram.com/NegativeBinomialDistribution.html) and The R Stats Package (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/family.html).

Dan Thompson is in the Division of Family Health Services, Bureau of Family and Community Health. He also serves as adjunct faculty at Florida State University. Note that the R programs used for this analysis are available from: Daniel Thompson, at email: dan_thompson@doh.state.fl.us. Email him at: dan_thompson@doh.state.fl.us.

The ordinary least squares (OLS) regression method assumes the numerators and denominators follow a Gaussian or Normal distribution and the dependent variable is the natural logarithm of the rates. This method is recommended by the National Cancer Institute and has been used by Olson, et al. to analyze trends in incidence of Primary Central Nervous System Non-Hodgkin Lymphoma in the U.S. (Olson, et al., 2002).

When these methods are applied to randomly generated data, the probability of observing a statistically significant result should be close to the alpha level selected for the test. This is usually 0.05. The performance of these methods can be assessed by repeatedly applying them to randomly generated data and calculating the proportion of trials that result in statistical significance. If the tests are performing well in situations where the null hypothesis is true and there is no trend, this proportion should be close to the alpha level. This is generally known as a Monte Carlo experiment.

Monte Carlo experiments can also be used to assess the performance of these methods when there is a trend and the null hypothesis of no trend is false. Ideally, a method that performs well would detect a trend, when the null hypothesis of no trend is true, in about 5% of the tests; and when the null hypothesis of no trend is false, the ideal method would detect a significant trend in a relatively high proportion of the tests, compared to the other methods. In this analysis, Monte Carlo experiments were used to evaluate and compare the four methods discussed above. The objective is to provide a better understanding regarding the choice of the appropriate method for a given situation.

Methodology

R software (The R Project for Statistical Computing available at: http://www.r-project.org/) was used to randomly generate 10 sets of numerators and denominators. These were then used to calculate simulated rates. Random data were generated based on means and standard deviations from four different sets of numerators and denominators taken from actual statistics for the period 1996 through 2005 (Florida Community Health Assessment Resource Tool Kit (CHARTS) at: http://www.floridacharts.com/charts/chart.aspx). The four data sets used were:

1) Injury mortality data for Florida;
2) Infant mortality (death before age 1) data for Florida;
3) Infant low birth weight (birth weight < 2500 grams) data for a Florida county; and
4) Infant mortality data for a Florida County.

The means and standard deviations for the numerators and denominators in these 4 data sets are given in table 1.

The data were generated to follow either a Normal (Gaussian) or a Poisson distribution. The 4 methods described in the Introduction were applied to the data sets and the results were compared. These methods were used to derive the equation that best fit the trend. The equation slope coefficient and the standard deviation of the slope coefficient were used to test for a statistically significant trend. The glm (generalized linear model) function in R was used to generate the equations.

This process was repeated 10,000 times and the proportion of trials that indicated statistical significance was compared for the 4 methods. In general, when statistical tests are applied to random data, where the null hypothesis is true, statistical significance will be observed in a proportion close to the alpha level of the test. This follows because the alpha level is defined as the probability of rejecting the null hypothesis when the null hypothesis is true. With trend data, the null hypothesis asserts there is no underlying trend and any observed trend is due to random variation. The four methods were compared in terms of their ability to accept the null hypothesis when the null hypothesis of no trend is true.

The four methods were also assessed for their ability to reject the null hypothesis of no trend when it is false. In this process the random data were generated as described above and then each succeeding simulated year of cases was increased by 1%. The formula for this simulated increase was $(1.01)^{(n-1)}$, where n is the year numbers 1 through 10. These data were generated for 10,000 simulated 10 year periods

Table 1: Means and Standard Deviations from Four Different Sets of Numerators and Denominators Taken From Florida Community Health Statistics 1996-2005

Means and Standard Deviations Used to Generate Simulated Data Sets

Data Set	Numerator Mean	Numerator Stand. Dev.	Denominator Mean	Denominator Stand. Dev.
Florida Injury Mortality	10,293.00	1,311.00	16,275,762.00	1,119,822.00
Florida Infant Mortality	1,483.00	87.80	205,609.00	11,707.00
Florida single county LBW	274.10	27.12	2,983.60	117.04
Florida single county infant mortality	27.50	5.82	2,983.60	117.04

and, as described above, the four methods were used to test for significant trends.

Results

The tables below give the results of the Monte Carlo trials. In the simulations where the null hypothesis of no trend was true, the statistical tests using OLS, Quasipoisson and Negative Binomial regression methods performed well when the data were normally distributed and also when the data followed a Poisson distribution.

As expected, with an alpha level of 0.05, approximately 2.5% of the trials reached statistically high significance and approximately 2.5% reached statistically low significance (Tables 2 through 5). In contrast, the Poisson regression method performed well only when the numerators and denominators followed the Poisson distribution. In simulations where the data followed a Normal distribution, and the null hypothesis of no trend was true, the Poisson regression method indicated statistical significance in far more than 5% of the simulations (Tables 2 through 5). The results for the Poisson method were better for the smaller data sets.

For example, in the test data set with the largest numerators and denominators (Table 2) the Poisson method indicated a significant trend in almost 90% (45.57% significantly low plus 44.24% significantly high) of the simulations where the null hypothesis of no trend was true, while the other 3 methods indicated a significant trend in a proportion close to the alpha level of 5%. The Poisson method performed better as the size of the numerators and denominators in the test data sets became smaller. For the data set with the smallest numerators and denominators (Table 5) the Poisson method indicated significance in 8.24% of the simulations where the null hypothesis was true, which is much closer to the desired alpha level of 5%.

In the results for the simulations where the null hypothesis of no trend was false (Tables 2 through 5), three of the four methods performed about equally, when the data were normally distributed. In contrast, the Poisson regression method detected the trend in larger proportions of the trails. For example, in Table 2 for the simulations with the normally distributed data, where the null hypothesis was false, the Poisson method detected a significant trend in 68.66% of the simulations. The other 3 methods all detected a significant trend in about 8% of the simulations.

Based on these data, it appears the Poisson method is more likely to detect a trend when the null hypothesis of no trend is false, but, as shown in tables 2 through 5, the Poisson method is also more likely to detect a trend when the null hypothesis of no trend is true. In short, with normally distributed data, the Poisson method is more likely to detect a trend when a trend is present and also when a trend is not present.

When the data followed a Poisson distribution, and the null hypothesis of no trend was false, the Poisson method was more likely to detect a significant tend compared to the other 3 methods. For example, in Table 3, in the simulations where the null hypothesis is false, the Poisson method detected a trend in 94.04% of the Poisson simulations, while the other 3

methods detected a significant trend in about 86% of the Poisson simulations. In contrast to the simulations of normally distributed data, the Poisson method was not more likely to detect a trend, when the null hypothesis of no trend was true, when the simulated data followed a Poisson distribution. In summary, the Poisson method performed as well as the other 3 methods when the data followed a Poisson distribution, and the null hypothesis of no trend was true. And the Poisson method was more likely to detect a trend when the null hypothesis of no trend was false and the simulated data followed a Poisson distribution.

Conclusion

These results indicate the Poisson regression method, for testing the statistical significance of rate trends, performs well only when the numerator and denominator data follow a Poisson distribution. The Ordinary Least Squares, Quasi-Poisson and Negative Binomial regression methods were more robust and performed well when the data were either Normally distributed or when they followed a Poisson distribution. When the simulation data followed a Poisson distribution and the null hypothesis of no trend was false, the Poisson regression method detected the trend more often

Table 2: Results of 10,000 Simulations of Florida Injury Mortality Rate Trends by Statistical Method and Distribution* Characteristics

Method	Test Data Distribution	Null Hypothesis: No Trend**	Percent Significantly Low	Percent Not Significant	Percent Significantly High
OLS Regression	Normal	TRUE	2.25%	95.20%	2.55%
Poisson Regression	Normal	TRUE	45.57%	10.19%	44.24%
Negative Binomial	Normal	TRUE	2.29%	95.11%	2.60%
Quasipoisson	Normal	TRUE	2.28%	95.17%	2.55%
OLS Regression	Poisson	TRUE	2.25%	95.27%	2.48%
Poisson Regression	Poisson	TRUE	2.33%	95.19%	2.48%
Negative Binomial	Poisson	TRUE	2.26%	95.26%	2.48%
Quasipoisson	Poisson	TRUE	2.27%	95.25%	2.48%
OLS Regression	Normal	FALSE	0.53%	91.60%	7.87%
Poisson Regression	Normal	FALSE	22.89%	8.45%	68.66%
Negative Binomial	Normal	FALSE	0.49%	91.71%	7.80%
Quasipoisson	Normal	FALSE	0.52%	91.73%	7.75%
OLS Regression	Poisson	FALSE	0.00%	0.00%	100.00%
Poisson Regression	Poisson	FALSE	0.00%	0.00%	100.00%
Negative Binomial	Poisson	FALSE	0.00%	0.00%	100.00%
Quasipoisson	Poisson	FALSE	0.00%	0.00%	100.00%

* Simulated 10 years of Florida injury mortaltiy rates with randomly generated numerators at mean 10,293 and denominators at mean 16,275,762. For the random normal data, the standard deviations were 1,311 for the numerators and 1,119,822 for the denominators. For the random Poisson data, the standard deviations were the square roots of the means.

** Where Null Hypothesis of no trend = FALSE, average trend = 0.01 increase per year

Table 3: Results of 10,000 Simulations of Florida Infant Mortality Rate Trends by Statistical Method and Distribution* Characteristics

Method	Test Data Distribution	Null Hypothesis: No Trend**	Percent Significantly Low	Percent Not Significant	Percent Significantly High
OLS Regression	Normal	TRUE	2.51%	94.88%	2.61%
Poisson Regression	Normal	TRUE	26.23%	46.62%	27.15%
Negative Binomial	Normal	TRUE	2.51%	94.88%	2.61%
Quasipoisson	Normal	TRUE	2.47%	94.89%	2.64%
OLS Regression	Poisson	TRUE	2.44%	94.92%	2.64%
Poisson Regression	Poisson	TRUE	2.26%	95.23%	2.51%
Negative Binomial	Poisson	TRUE	2.42%	94.93%	2.65%
Quasipoisson	Poisson	TRUE	2.43%	94.93%	2.64%
OLS Regression	Normal	FALSE	0.11%	83.87%	16.02%
Poisson Regression	Normal	FALSE	3.91%	26.50%	69.59%
Negative Binomial	Normal	FALSE	0.11%	83.87%	16.02%
Quasipoisson	Normal	FALSE	0.10%	83.98%	15.92%
OLS Regression	Poisson	FALSE	0.00%	14.48%	85.52%
Poisson Regression	Poisson	FALSE	0.00%	5.96%	94.04%
Negative Binomial	Poisson	FALSE	0.00%	14.44%	85.56%
Quasipoisson	Poisson	FALSE	0.00%	14.50%	85.50%

* Simulated 10 years of Florida infant death rates with randomly generated numerators at mean 1,483 and denominators at mean 204,609. For the random normal data, the standard deviations were 87.8 for the numerators and 11,707 for the denominators. For the random Poisson data, the standard deviations were the square roots of the means

** Where Null Hypothesis of no trend = FALSE, average trend = 0.01 increase per year

Table 4: Results of 10,000 Simulations of Low Birth Weight Rate Trends for a Florida County by Statistical Method and Distribution* Characteristics

Method	Test Data Distribution	Null Hypothesis: No Trend**	Percent Significantly Low	Percent Not Significant	Percent Significantly High
OLS Regression	Normal	TRUE	2.76%	95.09%	2.15%
Poisson Regression	Normal	TRUE	13.76%	73.39%	12.85%
Negative Binomial	Normal	TRUE	2.82%	95.01%	2.17%
Quasipoisson	Normal	TRUE	2.85%	94.95%	2.20%
OLS Regression	Poisson	TRUE	2.53%	95.02%	2.45%
Poisson Regression	Poisson	TRUE	2.92%	94.27%	2.81%
Negative Binomial	Poisson	TRUE	2.53%	95.02%	2.45%
Quasipoisson	Poisson	TRUE	2.53%	94.99%	2.48%
OLS Regression	Normal	FALSE	0.38%	88.83%	10.79%
Poisson Regression	Normal	FALSE	2.86%	57.22%	39.92%
Negative Binomial	Normal	FALSE	0.35%	88.68%	10.97%
Quasipoisson	Normal	FALSE	0.35%	88.74%	10.91%
OLS Regression	Poisson	FALSE	0.10%	75.72%	24.18%
Poisson Regression	Poisson	FALSE	0.05%	66.03%	33.92%
Negative Binomial	Poisson	FALSE	0.10%	75.71%	24.19%
Quasipoisson	Poisson	FALSE	0.11%	75.74%	24.15%

* Simulated 10 years of one Florida county low birth weight rates with randomly generated numerators at mean 274.1 and denominators at mean 2983.6. For the random normal data, the standard deviations were 27.12 for the numerators and 117.04 for the denominators. For the random Poisson data, the standard deviations were the square roots of the means.

** Where Null Hypothesis of no trend = FALSE, average trend = 0.01 increase per year

than the other three methods. When the test data followed a Poisson distribution and the null hypothesis of no trend was true, the Poisson regression method performed as well as the other three methods. However, in the simulations where the null hypothesis of no trend was true and the data followed a normal distribution, the Poisson regression method was far too likely to result in statistical significance, while the other three methods resulted in statistical significance in proportions close to the alpha level of 0.05. In summary, the Poisson method performed as well or better than the other methods when the simulated data followed a Poisson distribution but did not perform as well as the other methods when the simulated data followed a normal distribution.

One of the defining characteristics of the Poisson distribution is the mean is equal to the variance. In situations where the variance exceeds the mean (this is referred to as over-dispersion), Poisson regression will tend to underestimate the variance and thereby increase the probability that random results are deemed statistically significant.

Based on the results of this analysis, one recommendation is data should be examined to assess whether it follows a Poisson distribution, and the Poisson regression method should be used only when this condition is met. In practical terms, when using the Poisson regression method, the mean should be approximately equal to the variance. When this is not the case, it would probably be better to use

Table 5: Results of 10,000 Simulations of Infant Mortality Trends for a Florida County by Statistical Method and Distribution* Characteristics

Method	Test Data Distribution	Null Hypothesis: No Trend**	Percent Significantly Low	Percent Not Significant	Percent Significantly High
OLS Regression	Normal	TRUE	2.53%	95.17%	2.30%
Poisson Regression	Normal	TRUE	3.93%	91.76%	4.31%
Negative Binomial	Normal	TRUE	2.63%	95.05%	2.32%
Quasipoisson	Normal	TRUE	2.55%	95.14%	2.31%
OLS Regression	Poisson	TRUE	2.54%	95.31%	2.15%
Poisson Regression	Poisson	TRUE	2.67%	95.16%	2.17%
Negative Binomial	Poisson	TRUE	2.53%	95.34%	2.13%
Quasipoisson	Poisson	TRUE	2.57%	95.36%	2.07%
OLS Regression	Normal	FALSE	1.02%	93.54%	5.44%
Poisson Regression	Normal	FALSE	1.63%	88.25%	10.12%
Negative Binomial	Normal	FALSE	0.97%	93.60%	5.43%
Quasipoisson	Normal	FALSE	0.94%	93.69%	5.37%
OLS Regression	Poisson	FALSE	1.00%	93.39%	5.61%
Poisson Regression	Poisson	FALSE	0.88%	92.19%	6.93%
Negative Binomial	Poisson	FALSE	0.98%	93.45%	5.57%
Quasipoisson	Poisson	FALSE	0.92%	93.57%	5.51%

* Simulated 10 years of one Florida county infant mortaltiy rates with randomly generated numerators at mean 27.5 and denominators at mean 2983.6. For the random normal data, the standard deviations were 5.82 for the numerators and 117.04 for the denominators. For the random Poisson data, the standard deviations were the square roots of the means.

** Where Null Hypothesis of no trend = FALSE, average trend = 0.01 increase per year

the OLS, Quasi-Poisson, or Negative Binomial, regression methods or a nonparametric method

This analysis addressed only trends with 10 discrete points and the test data were generated with characteristics specific to Florida infant death, Low birth weight and injury mortality data. Using more or less points and data with different distribution characteristics could, and probably would, lead to different results and conclusions. The results and conclusions from this analysis apply only to Florida low birth weight, infant death and injury mortality data or data that are very similar. A general conclusion might be that different methods perform differently depending at least in part on the characteristics of the data to which they are applied. Further research is needed to reach a better understanding of the strengths and weaknesses of these methods in various situations.

References

Dunteman, G. H., & Ho, M-H. R. (2006). *An introduction to generalized linear models*. Thousand Oaks, CA: Sage Publications Inc.

Liu, L., Ikeda, K., & Yamori ,Y. (2001). Changes in stroke mortality rates for 1950 to 1997: A great slowdown of decline trend in Japan. *Stroke 2001*, *32*, 1745-1749. Available at: http://stroke.ahajournals.org/cgi/content/full/32/8/1745

Botha, et al. (2003) Breast cancer incidence and mortality trends in16 European countries. *European Journal of Cancer*, *39*, 1718-1729.

Lieb, et al. (2007). HIV/AIDS Diagnoses among blacks - Florida, 1999-2004. *MMWR*, February 2, 2007*, 56*(4), 69-73.

Olsen, J. E., et al. (2002). The continuing increase in the incidence of primary central nervous system non-Hodgkin lymphoma. *Cancer*, October 1, 2002, *95*(*7*), 1504-1510.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 384-395

1538 – 9472/09/$95.00

Level Robust Methods Based on the Least Squares Regression Estimator

Marie Ng
University of Hong Kong

Rand R. Wilcox
University of Southern California

Heteroscedastic consistent covariance matrix (HCCM) estimators provide ways for testing hypotheses about regression coefficients under heteroscedasticity. Recent studies have found that methods combining the HCCM-based test statistic with the wild bootstrap consistently perform better than non-bootstrap HCCM-based methods (Davidson & Flachaire, 2008; Flachaire, 2005; Godfrey, 2006). This finding is more closely examined by considering a broader range of situations which were not included in any of the previous studies. In addition, the latest version of HCCM, HC5 (Cribari-Neto, et al., 2007), is evaluated.

Key words: Heteroscedasticity, level robust methods, bootstrap.

Introduction

Consider the standard simple linear regression model

$$Y_i=\beta_0+\beta_1X_{i1}+\varepsilon_i, i=1, ..., n, \quad (1)$$

where β_0 and β_1 are unknown parameters and ε_i is the error term. When testing the hypothesis,

$$H_0: \beta_1 = 0 \quad (2)$$

the following assumptions are typically made:

1. $E(\varepsilon_i) = 0$.
2. $Var(\varepsilon_i) = \sigma^2$ (Homoscedasticity).
3. ε_i's are independent of X.
4. ε_i's are independent and identically distributed (i.i.d).

This article is concerned with testing (2) when assumption 2 is violated.

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ be the least squares estimate of $\beta = (\beta_0, \beta_1)$. When there is homoscedasticity (i.e., assumption 2 holds), an estimate of the squared standard error of $\hat{\beta}$ is $Var(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$, where $\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - 2)$ is the usual estimate of the assumed common variance; X is the design matrix containing an $n \times 1$ unit vector in the first column and X_{i1}'s in the second column.

Marie Ng is an Assistant Professor in the Faculty of Education. Email: marieng@uw.edu. Rand R. Wilcox is a Professor of Psychology. Email: rwilcox@usc.edu.

However, when heteroscedasticity occurs, the squared standard error based on such an estimator is no longer accurate (White, 1980). The result is that the usual test of (2) is not asymptotically correct. Specifically, using the classic *t*-test when assumptions are violated can result in poor control over the probability of a Type I error. One possible remedy is to test the assumption that the error term is homoscedastic before proceeding with the *t*-test. However, it is unclear when the power of such a test is adequate to detect heteroscedasticity.

One alternative is to use a robust method that performs reasonably well under homoscedasticity and at the same time is robust to heteroscedasticity and non-normality. Many methods have been proposed for dealing with heteroscedasticity. For example, a variance stabilizing transformation may be applied to the dependent variable (Weisberg, 1980) or a weighted regression with each observation weighted by the inverse of the standard deviation of the error term may be performed (Greene, 2003).

Although these methods provide efficient and unbiased estimates of the coefficients and standard error, they assume that heteroscedasticity has a known form. When heteroscedasticity is of an unknown form, the best approach to date, when testing (2), is to use a test statistic (e.g., quasi-t test) based on a heteroscedastic consistent covariance matrix (HCCM) estimator. Several versions of HCCM have been developed that provide a consistent

and an unbiased estimate of the variance of coefficients even under heteroscedasticity (White, 1980; MacKinnon & White, 1985).

Among all the HCCM estimators, HC4 was found to perform fairly well with small samples (Long & Ervin, 2000; Cribari-Neto, 2004). Recently, Cribari-Neto, et al. (2007) introduced a new version of HCCM (HC5) arguing that HC5 is better than HC4, particularly at handling high leverage points in X. However, in their simulations, they only focused on models with $\varepsilon \sim N(0, 1)$. Moreover, only a limited number of distributions of X and patterns of heteroscedasticity were considered. In this study, the performances of HC5-based and HC4-based quasi-t statistics were compared by looking at a broader range of situations.

Methods combining an HCCM-based test statistic with the wild bootstrap method perform considerably better than non-bootstrap asymptotic approaches (Davidson & Flachaire, 2008; Flachaire, 2005). In a recent study, Godfrey (2006) compared several non-bootstrap and wild bootstrap HCCM-based methods for testing multiple coefficients ($H_0: \beta_1 = \ldots = \beta_p = 0$). It was found that when testing at the $\alpha = 0.05$ level, the wild bootstrap methods generally provided better control over the probability of a Type I error than the non-bootstrap asymptotic methods. However, in the studies mentioned, the wild bootstrap and non-bootstrap methods were evaluated in a limited set of simulation scenarios.

In Godfrey's study, data were drawn from a data set in Greene (2003) and only two heteroscedastic conditions were considered. Here, more extensive simulations were performed to investigate the performance of the various bootstrap and non-bootstrap HCCM-based methods. More patterns of heteroscedasticity were considered, as well as more types of distributions for both X and ε. Small sample performance of one non-bootstrap and two wild bootstrap versions of HC5-based and HC4-based quasi-t methods were evaluated.

Finally, two variations of the wild bootstrap method were compared when generating bootstrap samples. One approach makes use of the lattice distribution. Another approach makes use of a standardized continuous uniform distribution: Uniform(-1, 1). The former approach has been widely considered (Liu, 1988; Davidson & Flachaire, 2000; Godfrey, 2006) and was found to work well in various multiple regression situations. Of interest is how these two approaches compare when testing (2) in simple regression models. Situations were identified where wild bootstrap methods were unsatisfactory.

Methodology

HC5-Based Quasi-T Test (HC5-T)

The HC5 quasi-t statistic is based on the standard error estimator HC5, which is given by

$$\ddot{V} = (X'X)^{-1}X'\text{diag}\left[\frac{r_i^2}{(\quad)\ \sqrt{(1-h_{ii})^{\alpha_i}}}\right]X(X'X)^{-1}, \tag{3}$$

where r_i, i = 1, ..., n are the usual residuals, X is the design matrix,

$$\alpha_i = \min\{\frac{h_{ii}}{\bar{h}}, \max\left\{4, \frac{kh_{max}}{\bar{h}}\right\}\}$$
$$= \min\{\frac{nh_{ii}}{\sum_{i=1}^{n} h_{ii}}, \max\left\{4, \frac{nkh_{max}}{\sum_{i=1}^{n} h_{ii}}\right\}\} \tag{4}$$

and

$$h_{ii} = x_i\left(X'X\right)^{-1}x_i' \tag{5}$$

where x_i is the i^{th} row of X, $h_{max} = \max\{h_{11}, \ldots, h_{nn}\}$ and k is set at 0.7 as suggested by Cribari-Neto, et al. (2007, 2008). The motivation behind HC5 is that when high leverage observations are present in X, the standard error of the coefficients are often underestimated. HC5 attempts to correct such a bias by taking into account the maximal leverage.

For testing (2), the quasi-t test statistic is,

$$T = \hat{\beta}_1 - 0 / \sqrt{\ddot{V}_{22}} \tag{6}$$

where $\ddot{V}_{22}$ is the 2nd entry along the diagonal of $\ddot{V}$. Reject (2) if $|T| \geq t_{1-\alpha/2}$ where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student's t-distribution with $n - 2$ degrees of freedom.

HC4-Based Quasi-T Test (HC4-T)

The HC4 quasi-t statistics is similar to HC5-T, except the standard error is estimated using HC4, which is given by

$$\tilde{V} = \left(X'X\right)^{-1} X'\text{diag}\left[\frac{r_i^2}{\left(1-h_{ii}\right)^{\delta_i}}\right] X\left(X'X\right)^{-1} \quad (7)$$

where

$$\delta_i = \min\{4, \frac{h_{ii}}{\bar{h}}\} = \min\{4, \frac{nh_{ii}}{\sum_{i=1}^{n} h_{ii}}\} \quad (8)$$

HC5-Based Wild Bootstrap Quasi-T Test (HC5WB-D and HC5WB-C)

The test statistic for testing (2) is computed using the following steps:

1. Compute the HC5 quasi-t test statistics (T) given by (6).
2. Construct a bootstrap sample $Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + a_i r_i$, i = 1, ..., n, where a_i is typically generated in one of two ways. The first generates a_i from a two-point (lattice) distribution:

$$a_i = \begin{cases} -1, & \text{with probability } 0.5 \\ 1, & \text{with probability } 0.5 \end{cases}$$

   The other method uses

$$a_i = \sqrt{12}(U_i - 0.5)$$

   where U_i is generated from a uniform distribution on the unit interval. We denote the method based on the first approach HC5WB-D and the method based on the latter approach HC5WB-C.

3. Compute the quasi-t test statistics (T^*) based on this bootstrap sample, yielding

$$T^* = \frac{\hat{\beta}_1^* - 0}{\sqrt{\ddot{V}_{22}^*}} \quad (9)$$

4. Repeat Steps 2 - 3 *B* times yielding T_b^* , b = 1, ..., *B*. In the current study, *B* = 599.
5. Finally, a *p*-value is computed:

$$p = \frac{\#\{|T_b^*| \geq |T|\}}{B} \quad (10)$$

6. Reject H_0 if $p \leq \alpha$.

HC4-Based Wild Bootstrap Quasi-T Test (HC4WB-D and HC4WB-C)

The procedure for testing (2) is the same as that of HC5WB-D and HC5W-C except that HC4 is used to estimate the standard error.

Simulation Design

Data are generated from the model:

$$Y_i = X_i\beta_1 + \tau(X_i)\varepsilon_i \quad (11)$$

where τ is a function of X_i used to model heteroscedasticity. Data are generated from a *g*-and-*h* distribution. Let Z be a random variable generated from a standard normal distribution,

$$X = \left(\frac{\exp(gZ)-1}{g}\right)\exp(hZ^2/2) \quad (12)$$

has a *g*-and-*h* distribution. When *g* = 0, this last equation is taken to be $X = Z\exp(hZ^2/2)$. When *g* = 0 and *h* = 0, X has a standard normal distribution. Skewness and heavy-tailedness of the *g*-and-*h* distributions are determined by the values of *g* and *h*, respectively. As the value of *g* increases, the distribution becomes more skewed. As the value of *h* increases, the distribution becomes more heavy-tailed. Four types of distributions are considered for X: standard normal (*g* = 0, *h* = 0), asymmetric light-tailed (*g* = 0.5, *h* = 0), symmetric heavy-tailed (*g* = 0, *h* = 0.5) and asymmetric heavy-tailed (*g* = 0.5, *h* = 0.5). The error term (ε_i) is also randomly generated based on one of these four *g*-and-*h* distributions. When *g*-and-*h* distributions are asymmetric (*g* = 0.5), the mean is not zero. Therefore, ε_i's generated from these distributions are re-centered to have a mean of zero.

Five choices for $\tau(X_i)$ are considered: $\tau(X_i) = 1$, $\tau(X_i) = \sqrt{|X_i|}$, $\tau(X_i) = 1 + \frac{2}{|X_i|+1}$, and $\tau(X_i) = |X_i + 1|$. These functions are denoted as variance patterns (VP), VP1, VP2, VP3, VP4 and VP5, respectively. Homoscedasticity is represented by $\tau(X_i) = 1$.

Moreover, $\tau(X_i) = \sqrt{|X_i|}$, $\tau(X_i) = 1 + \frac{2}{|X_i|+1}$, and $\tau(X_i) = |X_i + 1|$ represent a particular pattern of variability in Y_i based upon the value of X_i. All possible pairs of X_i and ε_i distributions are considered, resulting in a total of 16 sets of distributions. All five variance patterns are used for each set of distributions. Hence, a total of 80 simulated conditions are

considered. The estimated probability of a Type I error is based on 1,000 replications with a sample size of n = 20 and when testing at $\alpha = 0.05$ and $\alpha = 0.01$. According to Robey and Barcikowski (1992), 1,000 replications are sufficient from a power point of view. If the hypothesis that the actual Type I error rate is 0.05 is tested, and power should be 0.9 when testing at the 0.05 level and the true α value differs from 0.05 by 0.025, then 976 replications are required. The actual Type I error probability is estimated with $\hat{\alpha}$, the proportion of *p*-values less than or equal to 0.05 and 0.01.

Results

First, when testing at both $\alpha = 0.05$ and $\alpha = 0.01$, the performances of HC5-T and HC4-T are extremely similar in terms of control over the probability of a Type I error (See Tables 1 and 2). When testing at $\alpha = 0.05$, the average Type I error rate was 0.038 (SD = 0.022) for HC5-T and 0.040 (SD = 0.022) for HC4-T. When testing at $\alpha = 0.01$, the average Type I error rate was 0.015 (SD = 0.013) for HC5-T and 0.016 (SD = 0.013) for HC4-T.

Theoretically, when leverage points are likely to occur (i.e. when X is generated from a distribution with $h = 0.5$), HC5-T should perform better than HC4-T; however, as shown in Table 1, this is not the case. On the other hand, when leverage points are relatively unlikely (i.e., when X is generated from a distribution with $h = 0$), HC5-T and HC4-T should yield the same outcomes. As indicated by the results of this study, when X is normally distributed ($g = 0$ and $h = 0$), the actual Type I error rates resulting from the two methods are identical. However, when X has a skewed light-tailed distribution ($g = 0.5$ and $h = 0$), HC5-T and HC4-T do not always yield the same results. Focus was placed on a few situations where HC4-T is unsatisfactory, and we considered the extent it improves as the sample size increases. We considered sample sizes of 30, 50 and 100. As shown in Table 3, control over the probability of a Type I error does not improve markedly with increased sample sizes.

Second, with respect to the non-bootstrap and bootstrap methods, results suggest that the bootstrap methods are not necessarily superior to the non-bootstrap ones. As shown in Figures 1 and 4, when testing at $\alpha = 0.05$, under VP 1 and 4, the bootstrap methods outperform the non-bootstrap methods. Specifically, the non-bootstrap methods tended to be too conservative under those conditions. Nonetheless, under VP 3 and 5 (see Figures 3 and 5), the non-bootstrap methods, in general, performed better than the bootstrap methods. In particular, the actual Type I error rates yielded by the bootstrap methods in those situations tended to be noticeably higher than the nominal level. In one situation, the actual Type I error rate was as high as 0.196. When testing at $\alpha = 0.01$, HC5WB-C and HC4WB-C offered the best performance in general; however, situations were found where non-bootstrap methods outperform bootstrap methods.

Finally, regarding the use of the continuous uniform distribution versus the lattice distribution for generating bootstrap samples, results suggest that the former has slight practical advantages. When testing at $\alpha = 0.05$, the average Type I error rates yielded by the two approaches are 0.059 for HC5WB-C and HC4WB-C and 0.060 for HC5WB-D and HC4WBD. When testing at $\alpha = 0.01$, the average Type I error rates are 0.015 for HC5WB-C and HC4WB-C and 0.021 for HC5WB-D and HC4WB-D. Overall, the actual Type I error rates yielded by HC5WB-C and HC4WB-C appear to deviate from the nominal level in fewer cases.

Conclusion

This study expanded on extant simulations by considering ranges of non-normality and heteroscedasticity that had not been considered previously. The performance of the latest HCCM estimator (HC5) was also closely considered. The non-bootstrap HC5-based and HC4-based quasi-t methods (HC5-T and HC4-T) were compared, as well as their wild bootstrap counterparts (HC5WB-D, HC5WB-C, HC4WB-D and HC4WB-C). Furthermore, two wild bootstrap sampling schemes were evaluated - one based on the lattice distribution; the other based on the continuous standardized uniform distribution.

As opposed to the findings of Cribari-Neto, et al. (2007), results here suggest that HC5 does not offer striking advantages over HC4.

Both HC5-T and HC4-T perform similarly across all the situations considered. In many cases, HC5-T appears more conservative than HC4-T. One concern is that, for the situations at hand, setting k = 0.7 when calculating HC5 may not be ideal; thus, whether changing the value of k might improve the performance of HC5-T was examined. As suggested by Cribari-Neto, et al., values of k between 0.6 and 0.8 generally yielded desirable results, for this reason k = 0.6 and k = 0.8 were considered. However, as indicated in Tables 4 and 5, regardless of the value of k, no noticeable difference was identified between the methods.

Moreover, contrary to both Davidson and Flachaire (2008) and Godfrey's (2006) findings, when testing the hypothesis $H_0: \beta_1 = 0$ in a simple regression model, the wild bootstrap methods (HC5WB-D, HC5WB-C, HC4WB-D and HC4WB-C) do not always outperform the non-bootstrap methods (HC5-T and HC4-T). By considering a wider range of situations, specific circumstances where the non-bootstrap methods outperform the wild bootstrap methods are able to be identified and vice versa. In particular, the non-bootstrap and wild bootstrap approaches are each sensitive to different patterns of heteroscedasticity.

For example, the wild bootstrap methods generally performed better than the non-bootstrap methods under VP 1 and 4 whereas the non-bootstrap methods generally performed better than the wild bootstrap methods under VP 3 and 5. Situations also exist (1988), Davidson and Flachaire (2008) and Godfrey (2006). The actual Type I error rates resulting from the methods HC5WB-C and HC4WB-C were generally less variable compared to those resulting from HC5WB-D and HC4WB-D. In many cases, the performances between the two approaches are similar, but in certain situations such as in VP3, HC5WB-C and HC4WB-C notably outperformed HC5WB-D and HC4WB-D.

References

Cribari-Neto, F. (2004). Asymptotic inference under heteroscedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*, 215-233.

Cribari-Neto, F., Souza, T. C., & Vasconcellos, A. L. P. (2007). Inference under heteroskedasticity and leveraged data. *Communication in Statistics -Theory and Methods*, *36*, 1877-1888.

Cribari-Neto, F., Souza, T. C., & Vasconcellos, A. L. P. (2008). Errata: Inference under heteroskedasticity and leveraged data. *Communication in Statistics -Theory and Methods*, *36*, 1877-1888. *Communication in Statistics -Theory and Methods*, *37*, 3329-3330.

Davidson, R., & Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics, 146,* 162-169.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, *49*, 361-376.

Greene, W. H. (2003). *Econometric analysis* (*5th Ed.*). New Jersey: Prentice Hall.

Godfrey, L. G. (2006). Tests for regression models with heteroscedasticity of unknown form. *Computational Statistics & Data Analysis*, *50*, 2715-2733.

Liu, R. Y. (1988). Bootstrap procedures under some non i.i.d. models. *The Annuals of Statistics*, *16*, 1696-1708.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217-224.

MacKinnon, J. G., & White, H. (1985). Some heteroscedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305-325.

Robey, R. R. & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283–288.

Weisberg, S. (1980). *Applied Linear Regression*. NY: Wiley.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, *48*, 817-838.

Table 1: Actual Type I Error Rates when Testing at $\alpha = 0.05$

X		e		VC	HC5WB-C	HC5WB-D	HC5-T	HC4WB-C	HC4WB-D	HC4-T
g	h	g	h							
0	0	0	0	1	0.051	0.036	0.030	0.050	0.035	0.030
				2	0.077	0.059	0.068	0.075	0.061	0.068
				3	0.052	0.036	0.048	0.053	0.038	0.048
				4	0.058	0.040	0.038	0.055	0.042	0.038
				5	0.085	0.064	0.072	0.086	0.065	0.072
0	0	0	0.5	1	0.048	0.044	0.022	0.052	0.046	0.022
				2	0.053	0.048	0.032	0.055	0.050	0.032
				3	0.055	0.054	0.036	0.054	0.051	0.036
				4	0.058	0.046	0.022	0.053	0.046	0.022
				5	0.053	0.054	0.036	0.055	0.051	0.036
0	0	0.5	0	1	0.059	0.047	0.041	0.055	0.045	0.041
				2	0.066	0.063	0.046	0.063	0.059	0.046
				3	0.058	0.050	0.054	0.058	0.047	0.054
				4	0.065	0.054	0.038	0.066	0.056	0.038
				5	0.093	0.074	0.072	0.091	0.070	0.072
0	0	0.5	0.5	1	0.036	0.028	0.017	0.036	0.030	0.017
				2	0.044	0.037	0.024	0.045	0.040	0.024
				3	0.057	0.053	0.036	0.054	0.056	0.036
				4	0.048	0.045	0.018	0.051	0.046	0.018
				5	0.157	0.152	0.118	0.165	0.154	0.118
0	0.5	0	0	1	0.053	0.049	0.028	0.060	0.050	0.033
				2	0.059	0.050	0.043	0.063	0.056	0.049
				3	0.051	0.055	0.039	0.056	0.053	0.043
				4	0.048	0.035	0.018	0.042	0.030	0.020
				5	0.060	0.051	0.045	0.059	0.050	0.052
0	0.5	0	0.5	1	0.044	0.042	0.008	0.045	0.041	0.009
				2	0.055	0.063	0.024	0.050	0.063	0.028
				3	0.043	0.054	0.023	0.042	0.048	0.027
				4	0.036	0.038	0.007	0.032	0.035	0.008
				5	0.043	0.068	0.029	0.044	0.067	0.031
0	0.5	0.5	0	1	0.058	0.044	0.031	0.051	0.048	0.033
				2	0.070	0.054	0.053	0.067	0.058	0.056
				3	0.054	0.052	0.047	0.056	0.051	0.050
				4	0.050	0.041	0.023	0.050	0.039	0.024
				5	0.055	0.052	0.041	0.056	0.055	0.045
0	0.5	0.5	0.5	1	0.039	0.043	0.013	0.042	0.037	0.013
				2	0.048	0.055	0.026	0.043	0.053	0.030
				3	0.049	0.062	0.030	0.045	0.063	0.037
				4	0.023	0.042	0.006	0.024	0.046	0.006
				5	0.071	0.090	0.045	0.078	0.086	0.054
0.5	0	0	0	1	0.067	0.061	0.049	0.068	0.055	0.050
				2	0.070	0.057	0.055	0.068	0.057	0.060
				3	0.061	0.064	0.057	0.064	0.064	0.058
				4	0.061	0.048	0.038	0.066	0.047	0.038
				5	0.075	0.095	0.066	0.083	0.088	0.069

Table 1: Actual Type I Error Rates when Testing at $\alpha = 0.05$ (continued)

X		e		VC	HC5WB-C	HC5WB-D	HC5-T	HC4WB-C	HC4WB-D	HC4-T
g	h	g	h							
0.5	0	0	0.5	1	0.052	0.047	0.023	0.056	0.045	0.023
				2	0.056	0.059	0.029	0.053	0.055	0.031
				3	0.057	0.071	0.034	0.056	0.071	0.035
				4	0.041	0.036	0.020	0.043	0.041	0.020
				5	0.065	0.089	0.037	0.067	0.092	0.038
0.5	0	0.5	0	1	0.053	0.048	0.040	0.058	0.049	0.041
				2	0.073	0.055	0.072	0.081	0.064	0.073
				3	0.078	0.073	0.062	0.072	0.074	0.064
				4	0.046	0.038	0.027	0.040	0.040	0.027
				5	0.107	0.113	0.087	0.108	0.111	0.087
0.5	0	0.5	0.5	1	0.044	0.044	0.019	0.046	0.047	0.019
				2	0.065	0.062	0.050	0.068	0.065	0.051
				3	0.059	0.081	0.055	0.070	0.083	0.055
				4	0.048	0.046	0.019	0.046	0.048	0.019
				5	0.168	0.190	0.120	0.168	0.196	0.124
0.5	0.5	0	0	1	0.080	0.056	0.034	0.076	0.056	0.041
				2	0.062	0.065	0.040	0.064	0.067	0.047
				3	0.064	0.080	0.047	0.063	0.072	0.051
				4	0.050	0.042	0.017	0.047	0.038	0.019
				5	0.069	0.089	0.044	0.073	0.092	0.057
0.5	0.5	0	0.5	1	0.035	0.048	0.013	0.035	0.044	0.013
				2	0.038	0.057	0.017	0.036	0.059	0.018
				3	0.042	0.077	0.028	0.041	0.079	0.034
				4	0.036	0.036	0.007	0.028	0.033	0.008
				5	0.082	0.122	0.053	0.080	0.118	0.058
0.5	0.5	0.5	0	1	0.058	0.041	0.026	0.058	0.040	0.029
				2	0.061	0.057	0.043	0.061	0.055	0.054
				3	0.048	0.062	0.036	0.050	0.066	0.043
				4	0.045	0.038	0.016	0.049	0.035	0.016
				5	0.059	0.083	0.035	0.057	0.078	0.041
0.5	0.5	0.5	0.5	1	0.036	0.039	0.010	0.038	0.041	0.012
				2	0.057	0.057	0.021	0.055	0.059	0.031
				3	0.062	0.094	0.046	0.063	0.094	0.050
				4	0.030	0.041	0.007	0.036	0.036	0.008
				5	0.084	0.116	0.058	0.086	0.117	0.065
				Max	0.168	0.190	0.120	0.168	0.196	0.124
				Min	0.023	0.028	0.006	0.024	0.030	0.006
				Average	0.059	0.060	0.038	0.059	0.060	0.040
				SD	0.022	0.026	0.022	0.023	0.027	0.022

Table 2: Actual Type I Error Rates when Testing at $\alpha = 0.01$

X		e		VC	HC5WB-C	HC5WB-D	HC5-T	HC4WB-C	HC4WB-D	HC4-T
g	h	g	h							
0	0	0	0	1	0.016	0.005	0.013	0.017	0.011	0.013
				2	0.016	0.009	0.016	0.014	0.009	0.016
				3	0.018	0.010	0.017	0.017	0.009	0.017
				4	0.016	0.010	0.008	0.017	0.011	0.008
				5	0.024	0.018	0.024	0.025	0.021	0.024
0	0	0	0.5	1	0.013	0.010	0.008	0.012	0.007	0.008
				2	0.013	0.018	0.010	0.013	0.015	0.010
				3	0.014	0.025	0.012	0.011	0.019	0.012
				4	0.006	0.004	0.001	0.008	0.005	0.001
				5	0.013	0.024	0.009	0.013	0.019	0.009
0	0	0.5	0	1	0.016	0.016	0.010	0.019	0.016	0.010
				2	0.022	0.011	0.023	0.021	0.011	0.023
				3	0.021	0.010	0.020	0.017	0.008	0.020
				4	0.012	0.011	0.011	0.013	0.010	0.011
				5	0.026	0.019	0.025	0.026	0.018	0.025
0	0	0.5	0.5	1	0.006	0.009	0.004	0.007	0.009	0.004
				2	0.015	0.010	0.010	0.013	0.010	0.010
				3	0.014	0.012	0.014	0.015	0.010	0.014
				4	0.008	0.010	0.001	0.006	0.008	0.001
				5	0.060	0.063	0.047	0.054	0.071	0.047
0	0.5	0	0	1	0.011	0.010	0.006	0.010	0.010	0.006
				2	0.010	0.007	0.015	0.013	0.008	0.018
				3	0.012	0.017	0.014	0.016	0.014	0.015
				4	0.005	0.002	0.003	0.006	0.004	0.004
				5	0.017	0.022	0.021	0.018	0.030	0.023
0	0.5	0	0.5	1	0.005	0.013	0.004	0.004	0.009	0.004
				2	0.005	0.019	0.009	0.008	0.022	0.011
				3	0.006	0.028	0.006	0.006	0.028	0.007
				4	0.007	0.007	0.004	0.006	0.006	0.004
				5	0.009	0.021	0.009	0.007	0.020	0.012
0	0.5	0.5	0	1	0.009	0.005	0.009	0.012	0.007	0.010
				2	0.016	0.020	0.020	0.018	0.016	0.023
				3	0.014	0.022	0.023	0.017	0.022	0.024
				4	0.005	0.007	0.006	0.006	0.006	0.006
				5	0.009	0.016	0.013	0.008	0.015	0.015
0	0.5	0.5	0.5	1	0	0.011	0	0.001	0.006	0
				2	0.009	0.018	0.010	0.007	0.016	0.012
				3	0.016	0.027	0.020	0.011	0.026	0.024
				4	0.004	0.012	0.001	0.004	0.011	0.001
				5	0.015	0.036	0.021	0.018	0.033	0.024
0.5	0	0	0	1	0.011	0.008	0.009	0.007	0.007	0.010
				2	0.019	0.021	0.023	0.021	0.021	0.027
				3	0.024	0.027	0.028	0.025	0.023	0.029
				4	0.015	0.008	0.009	0.012	0.009	0.009
				5	0.023	0.028	0.030	0.021	0.029	0.030

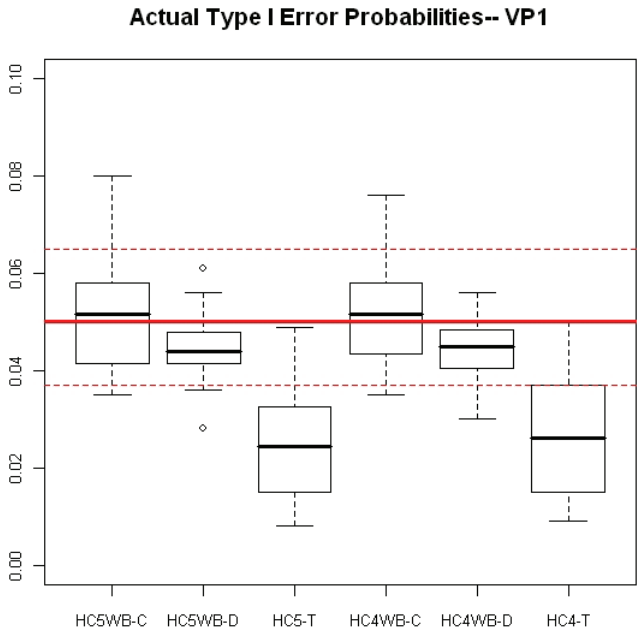
Table 2: Actual Type I Error Rates when Testing at α = 0.01 (continued)

X		e		VC	HC5WB-C	HC5WB-D	HC5-T	HC4WB-C	HC4WB-D	HC4-T
g	h	g	h							
0.5	0	0	0.5	1	0.008	0.007	0.005	0.007	0.004	0.005
				2	0.015	0.027	0.013	0.012	0.024	0.013
				3	0.013	0.031	0.009	0.014	0.030	0.009
				4	0.010	0.011	0.006	0.013	0.012	0.006
				5	0.021	0.057	0.015	0.023	0.057	0.015
0.5	0	0.5	0	1	0.010	0.007	0.005	0.010	0.010	0.005
				2	0.017	0.011	0.022	0.017	0.008	0.023
				3	0.026	0.034	0.040	0.029	0.031	0.040
				4	0.005	0.009	0.005	0.003	0.005	0.005
				5	0.028	0.049	0.030	0.023	0.05	0.031
0.5	0	0.5	0.5	1	0.010	0.011	0.004	0.008	0.011	0.004
				2	0.026	0.028	0.019	0.028	0.029	0.020
				3	0.032	0.039	0.032	0.033	0.041	0.034
				4	0.008	0.014	0.005	0.010	0.013	0.005
				5	0.078	0.114	0.072	0.076	0.111	0.079
0.5	0.5	0	0	1	0.008	0.005	0.011	0.011	0.005	0.012
				2	0.019	0.016	0.022	0.021	0.020	0.025
				3	0.007	0.036	0.013	0.006	0.037	0.013
				4	0.004	0.006	0.003	0.005	0.004	0.003
				5	0.012	0.034	0.014	0.012	0.032	0.021
0.5	0.5	0	0.5	1	0.004	0.011	0.002	0.006	0.011	0.002
				2	0.009	0.031	0.008	0.010	0.029	0.013
				3	0.006	0.029	0.008	0.008	0.027	0.010
				4	0.004	0.005	0	0.007	0.004	0.001
				5	0.074	0.114	0.075	0.081	0.116	0.076
0.5	0.5	0.5	0	1	0.003	0.003	0.006	0.005	0.004	0.006
				2	0.012	0.016	0.021	0.015	0.015	0.026
				3	0.015	0.027	0.022	0.016	0.030	0.026
				4	0.004	0.003	0	0.001	0.006	0
				5	0.017	0.036	0.020	0.014	0.038	0.024
0.5	0.5	0.5	0.5	1	0.010	0.011	0.004	0.010	0.014	0.004
				2	0.010	0.023	0.015	0.012	0.020	0.017
				3	0.014	0.045	0.021	0.013	0.047	0.029
				4	0.008	0.014	0.002	0.005	0.011	0.004
				5	0.025	0.059	0.024	0.027	0.060	0.031
				Max	0.078	0.114	0.075	0.081	0.116	0.079
				Min	0	0.002	0	0.001	0.004	0
				Average	0.015	0.021	0.015	0.015	0.021	0.016
				SD	0.013	0.020	0.013	0.013	0.020	0.014

Table 3: Actual Type I Error Rates when Testing at α = 0.05 with Sample Sizes 30, 50 and 100 for HC4-T

X		e		VP	n = 30	n = 50	n = 100
g	h	g	h				
0	0.5	0	0.5	1	0.022	0.021	0.018
0.5	0.5	0.5	0.5	1	0.012	0.019	0.020
0	0.5	0	0.5	4	0.014	0.007	0.011
0	0.5	0.5	0.5	4	0.011	0.009	0.023
0	0	0.5	0.5	5	0.118	0.123	0.143
0.5	0	0.5	0	5	0.093	0.070	0.078
0.5	0	0.5	0.5	5	0.190	0.181	0.174

Figure 1: Actual Type I Error Rates for VP1 when Testing at α = 0.05



The solid horizontal line indicates α = 0.05, the dashed lines indicate the upper and lower confidence limits for α, (0.037, 0.065).

Figure 2: Actual Type I Error Rates Under VP2

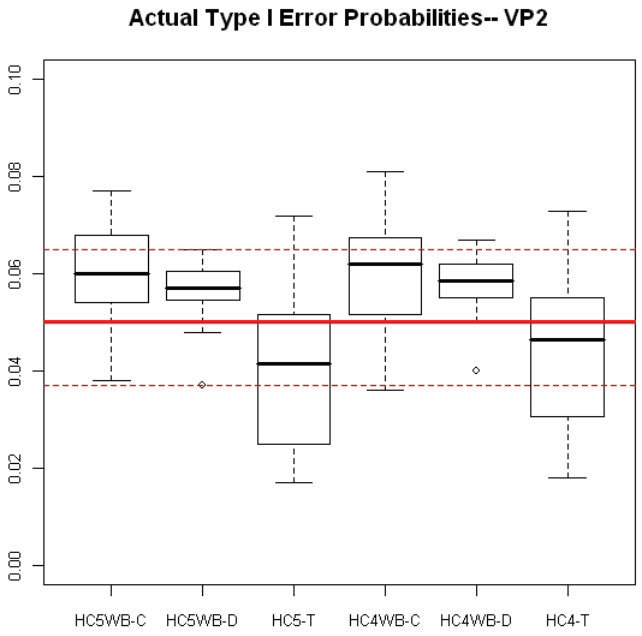


Figure 3: Actual Type I Error Rates Under VP3

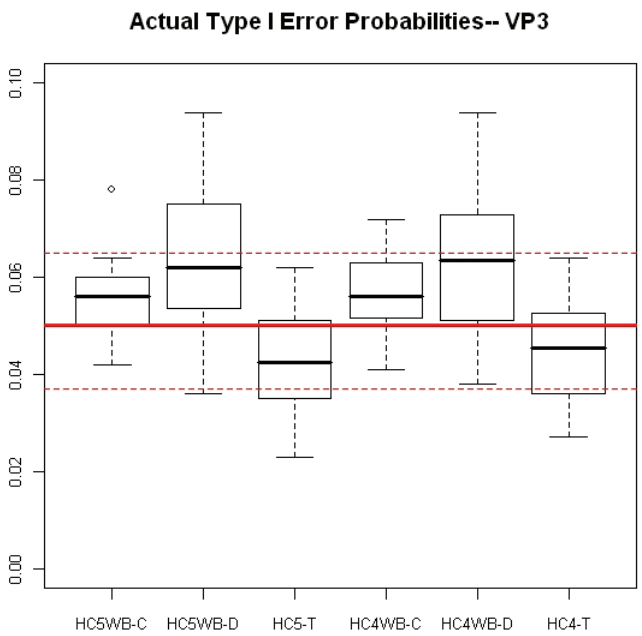


Figure 4: Actual Type I error rates under VP4

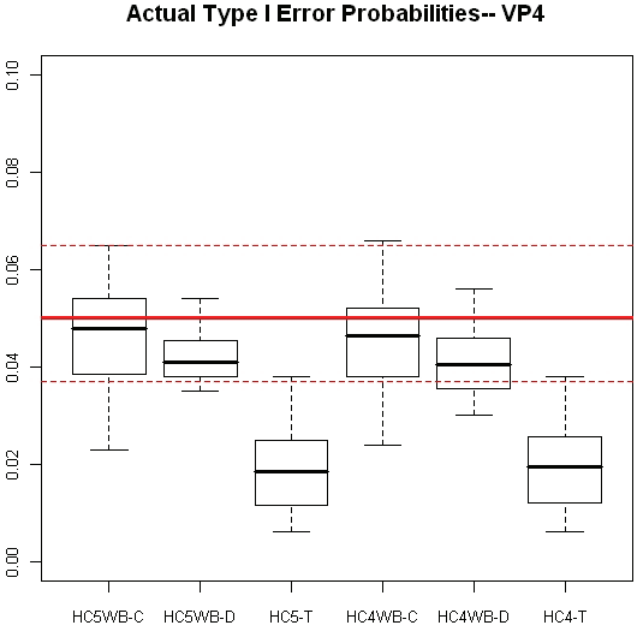


Figure 5: Actual Type I error rates under VP5

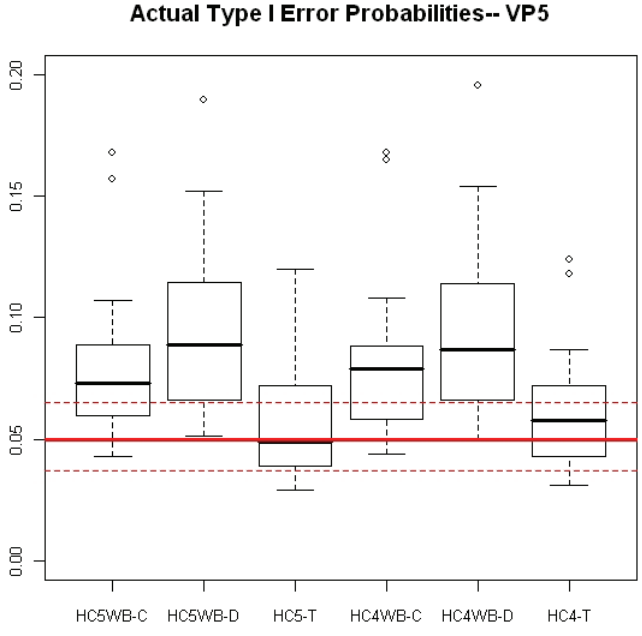


Table 4: Actual Type I Error Rates when Testing at α = 0.05, k = 0.6 for HC5-T

X		e		VP	HC5-T	HC4-T
g	h	g	h			
0	0.5	0.5	0.5	4	0.013	0.013
0.5	0.5	0	0.5	4	0.015	0.015
0	0	0.5	0.5	5	0.106	0.106
0.5	0	0.5	0.5	5	0.135	0.136

Table 5: Actual Type I Error Rates when Testing at α = 0.05, k = 0.8 for HC5-T

X		e		VP	HC5-T	HC4-T
g	h	g	h			
0	0.5	0.5	0.5	4	0.005	0.006
0.5	0.5	0	0.5	4	0.007	0.007
0	0	0.5	0.5	5	0.106	0.106
0.5	0	0.5	0.5	5	0.128	0.131

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 396-408

1538 – 9472/09/$95.00

Least Error Sample Distribution Function

Vassili F. Pastushenko
Johannes Kepler University of Linz, Austria

The empirical distribution function (ecdf) is unbiased in the usual sense, but shows certain order bias. Pyke suggested discrete ecdf using expectations of order statistics. Piecewise constant optimal ecdf saves 200%/N of sample size N. Results are compared with linear interpolation for U(0, 1), which require up to sixfold shorter samples at the same accuracy.

Key words: Unbiased, order statistics, approximation, optimal.

Introduction

Natural sciences search for regularities in the chaos of real world events at different levels of complexity. As a rule, the regularities become apparent after statistical analysis of noisy data. This defines the fundamental role of statistical science, which collects causally connected facts for subsequent quantitative analysis. There are two kinds of probabilistic interface between statistical analysis and empirical observations. In differential form this corresponds to histograms, and in integral form to the so-called sample distribution function or empirical distribution function (edf or ecdf in Matlab notation), c.f. Pugachev (1984), Feller (1971), Press, et al. (1992), Abramowitz & Stegun (1970), Cramér (1971), Gibbons & Chakraborti (2003). If histogram bins contain sufficiently big numbers of points, the usual concept of ecdf is more or less satisfactory. The focus of this paper is on short samples, where a histogram approach is not possible and an optimal integral approach is welcome. Consider i.i.d. sample X with N elements, numbered according to their appearance on the x-axis

$$X= [X_1, X_2,\ldots,X_N], \quad (1)$$

$$X_1 \le X_2 \le \ldots \le X_N. \quad (2)$$

Email Vassili F. Pastushenko at vassili.pastushenko@jku.at.

Sorted X-values are sometimes denoted $X_{(n)}$, but here parentheses are omitted. Parent d.f. F(x) is connected with corresponding p.d.f. $f(x)$

$$F(x) = \int_{-\infty}^{x} f(x)dx \quad (3)$$

F(x) is defined for the whole range of possible sample values between extreme x-values X_0 and X_{N+1} (denoted similarly to X for formal convenience):

$$X_0 = \inf(x),\ X_{N+1} = \sup(x) \quad (4)$$

Due to the fact that $f(x) \ge 0$, F(x) is non-decreasing. Therefore the exact random values of F(X), usually unknown in practical ecdf applications, are ordered according to positions of X elements at x-axis,

$$F_1 \le F_2 \le \ldots \le F_N. \quad (5)$$

where $F_1 = F(X_1)$, $F_2 = F(X_2)$, …, $F_N = F(X_N)$. For this reason values (5) are called order statistics, Gibbons & Chakraborti (2003). In literature ecdf is frequently denoted as $F_n(x)$ meaning that a sample consists of n elements. Here the notations are different. As defined in (5), $F_n = F(X_n)$, n = 1:N (colon is a convenient notation of MathWorks, meaning arithmetic progression between delimited expressions, here with an increment 1, more generally start : increment : finish). Usually ecdf is denoted $F_*(x, X)$, where x is the independent variable,

sometimes called parameter, taking any value of the principally possible X-values

$$F_*(x,X)=\frac{1}{N}\sum_{n=1}^{N}H(x-X_n) \qquad (6)$$

H(t) is Heaviside unit step function**,** H = 1 for t≥0, otherwise H=0. Function (6) as a piecewise constant approximation of F(x) takes N+1 values (levels) equal to (0:N)/N in N+1 x-intervals between and outside of N sample values. F_* is continuous from the right, although Pugachev (1984) suggested that the continuity from the left would be more reasonable. A centrally symmetrical version could be a compromise (H = 0.5 at t = 0). Middle points between adjacent F_* levels are

$$m = (n-0.5)/N, \quad n = 1:N \qquad (7)$$

For convenience, an example of F_* is shown for N = 3, Figure 1 A. Expected F_n-values (E_n, c.f. next section), shown by circles, are different from m.

Eq. (6) is constructed as an arithmetic mean of N ecdf, each corresponding to a 1-point sample,

$$F_*(x,X)=\frac{1}{N}\sum_{n=1}^{N}F_*(x,X_n) \qquad (8)$$

This shows that $E[F_*(x, X)] = F(x)$ for any N, where E[…] denotes the mathematical expectation, because this expectation corresponds to F_* for an infinitely long sample. In other words, for any fixed-in-advance x-value $F_*(x, X)$ represents an unbiased estimation of F(x). The name empirical reflects the similarity between $F_*(x, X)$, which gives the proportion of sample elements r satisfying r ≤ x, and F(x) = Prob(r≤x), r being a single random number. However, this similarity contains an arbitrary assumption. Indeed, differentiation of (8) with respect to x gives empirical p.d. f. $f_*(x, X)$, Feller (1971)

$$f_*(x)=\frac{1}{N}\sum_{n=1}^{N}\delta(x-X_n), \qquad (9)$$

$\delta(x)$ being the Dirac delta. As can be seen from this expression, ecdf (8) attributes probability measure 1/N to each sample element. As a result, any sample is represented as a set of measure 1, whereas in reality it represents a set of measure zero, which is obvious for the main class of continuous distributions, and discontinuous distributions can be considered as a limit of continuous ones. This contradiction is especially strongly expressed in eq.(6), where measure 1 is attributed to every single-point sample on the right-hand side, which should mean that every sample element is a deterministic, not a stochastic item.

As indicated by Pyke (1959), a more reasonable approach should consider a sample as a set of measure zero, which delimits N+1 nonzero-measure intervals on the x-axis. This is consistent with the point of view that the sampling procedure represents mapping of N random values of parent F(x) to the x-axis. A single random F-value is uniformly distributed in (0, 1), i.e., F∈ U(0, 1) . Each of the F-values mapped into the sample values is selected independently. However, these values finally appear on the F-axis as an ordered sequence, so that the neighbouring elements of the sequence are no longer independent. Order statistics F_1, …, F_N have their own distributions. Therefore, an optimal ecdf must use this information. Probability densities for random u ∈ U(0,1), u = F_n, are c.f. Gibbons & Chakraborti (2003), Durbin (1973), Pyke (1959):

$$f_{N,n}(u)=u^{n-1}(1-u)^{N-n}\frac{N!}{(n-1)!(N-n)!}, \quad n = 1:N. \qquad (10)$$

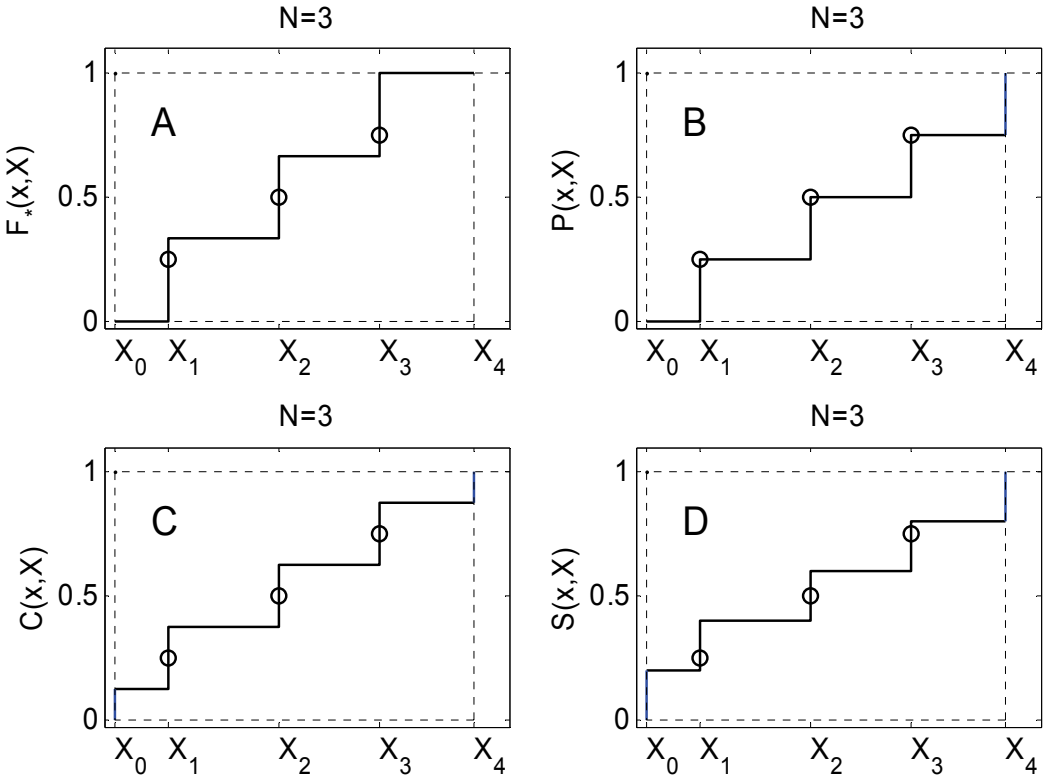
The first two moments of these distributions, or expected values and variances of F_n, denoted E_n and V_n respectively, are (c.f. Gibbons & Chakraborti (2003)):

$$E_n=E[F_n]=\int_0^1 x f_{N,n}(x)dx=\frac{n}{N+1}, \quad n=1:N \qquad (11)$$

and

$$V_n=\int_0^1 (x-E_n)^2 f_{N,n}(x)dx=\frac{n(N+1-n)}{(N+1)^2(N+2)} =\frac{E_n(1-E_n)}{N+2}, \quad n=1:N \qquad (12)$$

Figure 1: Different Ecdf-Versions for a Sample with 3 Elements
Expectations of order statistics are shown by circles. A: $F_*(x, X)$; B: P(x, X); C: C(x, X); D: S(x, X). Note that the numbers of jumps are different, A: N; B: N+1; C: N+2; D: N+2.



As a research tool, F_* is expected to optimally reproduce parent F(x). However, there are some discrepancies with predictions of order statistics (11-12). It follows from (6) that $E[F_*(X_n, X)] = E[n/N] = n/N$, n=1:N, whereas the correct expectation (11) is different , Pyke (1959). This discrepancy means a certain order bias.

Pyke (1959) considered order statistics F_n as zero-measure delimiters of probability intervals, created by the sample. He also considered statistic $C_N^+ = E_n - F_n$ instead of the usual statistic, $D_N^+ = n/N - F_n$, n = 1:N. This was interpreted by Brunk (1962), Durbin (1968) and Durbin and Knott (1972) as a discrete modification of ecdf. In particular, Brunk mentions Pyke's (1959) suggestion that the plotted points (F_n, n/(N+1)) in the Cartesian plane replace the empirical distribution function. In fact, as Hyndman and Fan (1996) mentioned, similar suggestions were made much earlier by Weibull (1939) and Gumbel (1939). These suggestions were partly considered for applications using ecdf values only at x = X, such as two-sided Kolmogorov-Smirnov test, Durbin (1968). However, any generalization for arbitrary x-values was not presented, although Hyndman and Fan (1996) discuss similar ideas concerning distribution quantiles.

To find an alternative for ecdf, an optimality criterion must be selected. Because the aim of ecdf is to approximate F(x), the main criterion is the approximation accuracy. Additionally convenience or simplicity may be discussed, but these aspects are almost the same within the class of piecewise constant approximations which are considered here.

The approximation accuracy needs a definition of distance between two compared distribution functions. The distance between two distributions, e.g. between exact F(x) and its empirical approximation, is frequently

characterized by the biggest (supremum) absolute value of their difference. Another possible measure could be an average absolute value of the difference. A more commonly used statistical measure is mean squared deviation, calculated either as a sum in discrete approaches or as an integral in continual approaches. For discrete applications, which only need ecdf values for the sample elements, Pyke's approach already gives an optimal solution in the sense of minimal rms deviation. Indeed, if it is desirable to replace a random F_n by a single number with minimal rms error, this number is E_n.

However, some applications need an extension of Pyke's discrete function to the whole range of possible x-values. Interpolation based on known knots (X_n, E_n) is one option. Linear interpolation, which was probably meant by Brunk's suggestion to plot (X_n, E_n), may work well numerically in some cases, such as uniform parent distribution, however, it is badly suited for unlimited distributions and it is difficult to obtain general, distribution-independent results. This article focuses on nearest interpolation, for which two versions are possible depending on the choice of an independent variable. A more attractive version corresponds to independent F. In this way, interpolating from known knots (E_n, X_n) to arbitrary (C, x), ecdf C(x, X) (Figure 1C) with expected E_n (circles) in the centres of corresponding probability intervals may be obtained. Table 1 lists the various notations used throughout this article.

Methodology

A family of sample distribution functions. An ecdf-like function P(x, X) can be constructed using (X_n, E_n):

$$P(x,X)=\frac{1}{N+1}\sum_{n=1}^{N}H(x-X_n). \qquad (13)$$

This function exactly corresponds to Pyke's suggestion at x = X. Nevertheless, P(x, X) is not very useful for arbitrary x-values, because it corresponds to a one-directional near interpolation, extending the optimal values only to the right. This is illustrated by Figure 1, B (E_n are shown by circles).

Vectors E = [E_1, …, E_n], and X = [X_1 , …, X_n], n=1:N, can be complimented by extremal values of F_n and x in order to enable interpolation in the entire range of x- and F-values. This leads to extended vectors **E** and **X**, each of size N+2:

$$\mathbf{E} = [0, E, 1] \qquad (14)$$

$$\mathbf{X}=[X_0, X, X_{N+1}] \qquad (15)$$

Two versions of the nearest interpolation are possible. In MathWorks syntax:

$$C= \text{interp}(\mathbf{X}, \mathbf{E}, x, \text{'nearest'}),\ X_0 \le x \le X_{N+1} \qquad (16)$$

and

$$x = \text{interp}(\mathbf{E}, \mathbf{X}, C, \text{'nearest'}),\ 0 \le C \le 1. \qquad (17)$$

Version (17) is more attractive for two reasons. First, **E** has known boundaries 0 and 1, whereas X_0 and/or X_{N+1} can be either unknown or infinite. Second, eq. (16) is less convenient for analysis because it involves middle points $m_{xn}=(\mathbf{X}_n+\mathbf{X}_{n-1})/2$, n=1:N+1, where any exact calculation of $E[F(m_{xn})]$ and $E[F(m_{xn})^2]$ for an unknown F(x) is not possible. As follows from (17),

$$C(x,X) = \frac{1}{N+1}\sum_{n=0}^{N+1}w_n H(x-\mathbf{X}_n) \qquad (18)$$

Weight coefficients w_n are equal to 1 except for n = 0 and n = N + 1, where w_n = 0.5. Thus eq. (18) attributes probability measure of 0.5/(N+1) to x-values below X_1 and above X_N respectively, formally to extremal x-values X_0 and X_{N+1}, and measure of 1/(N+1) to every sample element. Altogether, measure of N/(N+1) < 1 is now attributed to the very sample. Incomplete measure does not lead to any difficulty, because sample estimations based on (18) should be considered as conditional ones, and therefore the result should be normalized by condition probability N/(N+1). Thus, estimation of expected value of some function t(x) results in a traditional answer, mean(t(X)):

$$\frac{N+1}{N}\int_{X_1-0}^{X_N+0} t(x)\frac{d\,C(x,X)}{dx}dx = \frac{1}{N}\sum_{n=1}^{N} t(X_n) \tag{19}$$

Because extremal x-values acquire a probability measure, the first and last summands can be simplified in (18), which results in an equivalent of C in the entire x-range:

$$C(x,X) = \frac{1}{N+1}\left(\frac{1}{2}+\sum_{n=1}^{N} H(x-X_n)\right),$$

$$X_0 < x < X_{N+1}. \tag{20}$$

Table 1: Notations

[A, B, …]	Concatenation of A, B, …, a set consisting of A, B, …
C(x, X)	Centred ecdf, E-values are in the centres of corresponding probability intervals
d	Defect of levels, sum of their squared deviations from the optimal (natural) levels
D(α)	Total expected squared Deviation of s(x, X, α) from F(x)
D_*, D_C, D_S, …	Total expected squared deviations for F*(x, X), C(x, X), S(x, X), …
E = n/(N+1)	n = 1:N vector of expected order statistics.
E_n	n^{th} element of E
E = [0, E_1, …, E_N, 1]	Vector E extended by extremal E-values
E[<abc>]	Mathematical expectation of an expression <abc>
F(x)	Parent d.f.
f(x)	p.d.f., f = dF/dx
F_*(x, X)	Presently accepted ecdf
f_*(x)	Empirical p.d.f., $f_* = dF_*(x)/dx$
$f_{N,n}(u)$	p.d.f. of n-th order statistic u ∈ U(0,1), n = 1:N
g_z	Gain, relative total squared deviation (in units of total deviation for F_*), $g_z = D_z/D_*$, z = C,S,…
H(t)	Heaviside unit step, H=1 if t ≥ 0, otherwise H = 0. In Matlab: H = t >= 0
M = mean(X)	Average of sample elements
N	Sample size (length of i.i.d. sample)
P(x, X) NF_*(x, X)/(N+1)	Pyke function
s(x, X, α)	Family of ecdf with parameter α, 0 ≤ α < 0.5
s_n	Levels of s(x, X, α), n = 1:N+1
S(x, X)	Optimal member of s-family, minimizing D(α)
u	Uniform random variable, 0 ≤ u ≤ 1
U(0, 1)	Standard uniform distribution, F(u) = u, 0 ≤ u ≤ 1
X = [X_1, X_2, …, X_N]	i.i.d. sample with parent d.f. F(x)
X = [X_0, X_1, X_2, …, X_N, X_{N+1}]	Extended sample X by adding extremal x-values, size(**X**)= N+2
x	A number ∈ (set of possible X-values)
α, β	Parameters of ecdf family s(x, X, α, β)
δ(x)	Dirac delta
δ_{xX}	Kronecker symbol. In Matlab: δ_{xX} = any(x == X)
Δ	The deviation of an ecdf from the parent d.f.
Φ(x, X)	Hybrid of S and P for both continual and discrete applications

An example of C(x, X) is shown in Figure 1,C. Note that eq. (20) follows from eq. (13) by adding 0.5/(N+1), and C(x, X) has expected values E_n in the middle of the corresponding probability intervals. Therefore if the centrally symmetric unit step H(t) is accepted, C(X_n, X) automatically gives expected value E_n.

Functions P(x, X) and C(x, X) represent linear transformations of F_*, therefore F_*, P and C could be considered as members of two-parametric ecdf family *s*:

$$s(x,X,\alpha)=\alpha+\beta F_*(x,X) \qquad (21)$$

Thus, $\alpha=0, \beta=1$ leads to $s = F_*(x, X)$; $\alpha=0, \beta = N/(N+1)$ gives $s = P(x, X)$ and $\alpha = 0.5/(N+1)$, $\beta = N/(N+1)$ gives $s = C(x, X)$. Levels of P(x, X) are not symmetrical with respect to probability centre 0.5, i.e. not invariant in transformation levels $\rightarrow$1-levels. Therefore, although P(x, X) has expected values at x = X, it cannot be considered as a real alternative to F_*. Excluding P(x, X) from the s-family, the number of parameters may be reduced by setting $\beta = 1-2\alpha$, which enables the automorphism levels $\rightarrow$ 1-levels. This leads to one-parametric s-family

$$s(x,X,\alpha)=\alpha+(1-2\alpha)F_*(x,X), \quad 0\le\alpha<0.5. \qquad (22)$$

Levels s_n of s(x, X, α) follow from the levels of F_*:

$$s_n=\alpha+(1-2\alpha)(n-1)/N, \text{ n=1:N+1}, \qquad (23)$$

where $\alpha = 0$ corresponds to $F_*(x, X)$, and $\alpha = 0.5/(N+1)$ to C(x, X). Consider the properties of s(x, X, α) in terms of order statistics and squared deviation of s(x, X, α) from F(x).

Mean Values of F(x) Between Adjacent X_n and Natural Levels

As noted above, the mapping of F(x) to sample X leads to certain order statistics predictions (11-12), therefore,

$$E[F_n^2]=E_n^2+V_n=\frac{(n+1)E_n}{N+2};\ \text{n=1:N} \qquad (24)$$

In order to see how the levels s_n (23) agree with these predictions, different ecdf versions must be compared with F(x) within intervals ($\mathbf{X}_{n-1}$, $\mathbf{X}_n$) numbered by n=1:N+1. Consider the integrals:

$$I_{F,n}=\int_{X_{n-1}}^{X_n} F(x)f(x)dx=\int_{F_{n-1}}^{F_n} FdF=\frac{F_n^2-F_{n-1}^2}{2};$$
$$\text{n=1:N+1} \qquad (25)$$

and

$$I_{s,n}=\int_{X_{n-1}}^{X_n} s(x,X,)f(x)dx=s_n(F_n-F_{n-1});$$
$$\text{n=1:N+1} \qquad (26)$$

Integrals (25-26) represent another kind of order statistics. Natural levels S_n can be found from a comparison of their mathematical expectations, that is, from $E[I_{s,n}] = E[I_{F,n}]$, where

$$E[I_{F,n}]=\frac{E_n}{N+2};\ \text{n=1:N+1} \qquad (27)$$

and

$$E[I_{s,n}]=\frac{s_n}{N+1};\ \text{n= 1:N+1}. \qquad (28)$$

Equality of (27) and (28) leads to natural levels:

$$S_n=\frac{n}{N+2};\ \text{n = 1:N+1}. \qquad (29)$$

The levels follow if the right hand sides of (25 and 26) are equated and divided by F_n-F_{n-1}. The mathematical expectations found lead to levels of C(x, X):

$$C_n=E\left[\frac{F_n^2-F_{n-1}^2}{2(F_n-F_{n-1})}\right]=E\left[\frac{F_n+F_{n-1}}{2}\right]=\frac{2n-1}{2(N+1)};$$
$$\text{n = 1:N+1} \qquad (30)$$

Comparing the levels of F_*, given by (n-1)/N, and C_n (30) with natural levels S_n (29), n = 1:N+1, both are smaller than S_n below 0.5 and bigger than S_n above 0.5. If the ratio of differences between these levels is constructed and the natural ones, this ratio (for nonzero

values) appears to be greater than 2 at any N (zeros happen at the median level 0.5 for even N):

$$\frac{(n-1)/N - n/(N+2)}{(n-0.5)/(N+1) - n/(N+2)} = 2 + \frac{2}{N} \quad (31)$$

Thus, the detailed comparison leads to a conclusion: both levels of $F_*(x, X)$ and of $C(x, X)$ show certain order bias, because in average these levels do not match the expected behaviour of integrals of F between order statistics F_n. They are insufficiently big below the sample median, and too big above it.

The defect d of $s(x, X, \alpha)$ is introduced as a sum of squared deviations of s_n from natural levels (29),

$$d = \sum_{1}^{N+1} (s_n - S_n)^2. \quad (32)$$

The defect of F_* is

$$d_* = \sum_{n=1}^{N+1} \left(\frac{n-1}{N} - \frac{n}{N+2}\right)^2 = \frac{N+1}{3N(N+2)}, \quad (33)$$

and the defect of C is

$$d_C = \sum_{n=1}^{N+1} \left(\frac{n-0.5}{N+1} - \frac{n}{N+2}\right)^2 = \frac{N}{12(N+1)(N+2)}. \quad (34)$$

In agreement with eq. (31), the ratio of these defects is:

$$\frac{d_*}{d_C} = 4\left(1 + \frac{1}{N}\right)^2 \quad (35)$$

Two conclusions can be made. First, although near interpolation seems to be attractive in the sense that in puts expected values E_n exactly in the middle between C- levels, it is still not yet optimal S(x, X), based on natural levels (29):

$$S(x, X) = s(x, X, 1/(N+2)). \quad (36)$$

Thus, the optimum should occur at:

$$\alpha = \frac{1}{N+2}. \quad (37)$$

Ecdf S(x, X) formally ascribes to every element of the extended sample **X** probability measure of 1/(N+2):

$$S(x,X) = \frac{1}{N+2} \sum_{n=0}^{N+1} H(x - \mathbf{X}_n). \quad (38)$$

Ecdf S(x, X) has zero defect d by definition. Similar to C(x, X), the expression for S may be simplified as:

$$S(x,X) = \frac{1}{N+2}\left(1 + \sum_{n=1}^{N} H(x - X_n)\right), \quad X_0 < x < X_{N+1}. \quad (39)$$

An illustration of S(x, X) for N = 3 is given by Figure 1, D.

Results

Function S(x, X) Minimizes the Expected Total Error of F(x) Approximation.

It can be shown that S(x, X) minimizes the error of F(x) approximation by calculating total squared deviation D of $s(x, X, \alpha)$ from F(x) and finding an optimal α as $\text{argmin}(D(\alpha))$, getting in this way again $\alpha = 1/(N+2)$ as the optimal value. Total expected approximation error, or expected squared deviation is

$$D(\alpha) = E\left[\int_{X_0}^{X_{N+1}} (F(x) - s(x,X,\alpha))^2 f(x) dx\right] \quad (40)$$

The optimality of S is confirmed by following theorem and proof.

Theorem

S(x, X) represents least error approximation of F(x) at the family $s(x, X, \alpha)$, because it minimizes the total squared approximation error (40).

Proof

Consider deviation Δ,

$$\Delta = F(x) - s(x, X, \alpha) \quad (41)$$

as a random quantity at every fixed x due to randomness of X. Mathematical expectation of

Δ, taking into account eq. (22) and E[F∗(x, X)] = F(x), is:

$$E[\Delta] = F(x) - (\alpha+(1-2\alpha)F(x)) = \alpha\,(2F(x)-1). \quad (42)$$

The goal is to find $D = E[\Delta^2]$, therefore the variance, var(Δ), is needed. This can be found using the variance of F∗(x, X), expressed as F(x)(1-F(x))/N, Gibbons and Chakraborti (2003). Because in (41) F(x) is a deterministic function, only the second term in (41) contributes to var(Δ):

$$V_\Delta = \mathrm{var}(\Delta) = (1-2\alpha)^2 F(x)(1-F(x))/N. \quad (43)$$

Therefore, the expected squared deviation is:

$$E[\Delta^2 = V_\Delta + E[\Delta]^2 = (1-2\alpha)^2 F(x)(1-F(x))/N + \alpha^2 (2F(x)-1)^2 \quad (44)$$

Substituting (44) into (40) leads to total expected squared deviation D

$$\begin{aligned} D(\alpha) &= \int_{X_0}^{X_{N+1}} E[(F(x)-s(x,X,\alpha))^2]f(x)dx \\ &= \int_{X_0}^{X_{N+1}} \left[\begin{array}{l} (1-2\alpha)^2 \dfrac{F(x)(1-F(x))}{N} \\ +\alpha^2(2F(x)-1)^2 \end{array} \right] f(x)dx \\ &= \int_0^1 \left[(1-2\alpha)^2 \frac{F(1-F)}{N} + \alpha^2(2F-1)^2 \right] dF \\ &= \frac{2(N+2)\alpha^2 - 4\alpha + 1}{6N}. \end{aligned} \quad (45)$$

Thus, D(α) is quadratic in α with minimum at α defined by (37), which proves the theorem.

Now consider expected squared deviations for three different α-values leading to F∗, C and S. For α = 0 eq. (45) yields known result for F∗,

$$D_* = D(0) = \int_0^1 \frac{F(1-F)}{N} dF = \frac{1}{6N}. \quad (46)$$

For C(x, X), i.e. for α=0.5/(N+1):

$$D_C = D\left(\frac{0.5}{N+1}\right) = \frac{2N+1}{12(N+1)^2}, \quad (47)$$

and correspondingly, for S(x, X),

$$D_S = D\left(\frac{1}{N+2}\right) = \frac{1}{6(N+2)}. \quad (48)$$

Parabolic dependency of D(α), eq. (45) is illustrated in Figure 2 for several N-values. The values of D for three ecdf versions, F∗, C and S (46- 48), are indicated by special markers.

Linear Interpolation for Uniformly Distributed Data

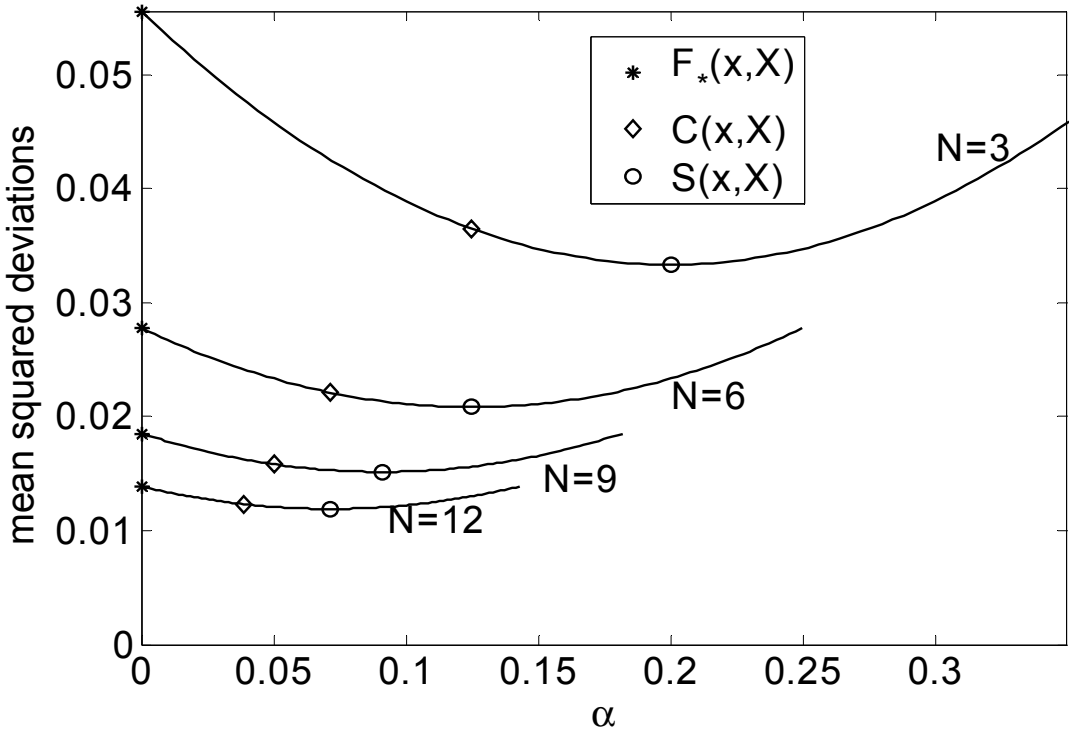
Compare the piecewise constant approximation in versions presented above with possibilities of different linear interpolations. In the case of a general parent d.f. F(x), it is difficult to get any analytical results. Therefore, F(x) is taken as standard uniform distribution, U(0, 1). However, this is more than a mere numerical example. Any known F(x) can be transformed to U(0, 1) by probability integral transformation u = F(x). Although in practice F(x) is mostly unknown, sometimes the transformation is possible, e.g. in fitting distribution parameters to X. Another meaningful aspect is - assuming that F(x) is known and transformed to standard uniform - the potentials of the linear interpolation become apparent.

Both versions of interpolation, eq. (16) and (17) are now considered linear instead of nearest. Let $E_{lin}(x, X)$ be ecdf, defined as interpolation between Pyke points (X_n, E_n) according to (16)

$$E_{lin}(x,X) = E_{n-1} + \frac{x - X_{n-1}}{X_n - X_{n-1}}(E_n - E_{n-1}); \quad X_{n-1} \le x \le X_n,$$
$$n = 1{:}N+1,\ X_{(1:N)} \in U(0, 1). \quad (49)$$

Here X in the left hand side is usual sample, and X in the right hand side is the extended sample, $X_0 = E_0 = 0$, $X_{N+1} = E_{N+1} = 1$.

Figure 2: Total Squared Expected Error D(α) of the Family s(x, X, α) for Several N-values
The cases of $F_*(x, X)$, C(x, X) and S(x, X) as members of s-family are shown by special symbols; note that min(D(α))-values (circles) linearly depend on optimal α-values.



Eq. (49) is nonlinear with respect to random numbers X. Correspondingly, the expectation $E[E_{lin}(x, X)]$ deviates from x. Expected squared deviations $E[(E_{lin}(x, X)-x)^2]$ were estimated numerically as an average over 10^5 X-samples at N = 5. Figure 3 compares the result with $E(\Delta^2_*)$, $E(\Delta^2_C)$ and $E(\Delta^2_S)$. The left figure shows these expectations for all four compared versions, and the right figure shows their integrals in (0, x), which give at x = 1 corresponding total errors D. The gains, shown on the top of the right figure, represent the relative total errors, i.e. D_C/D_*, D_S/D_* and D_{lin}/D_* respectively.

The total approximation error is notably smaller for linear interpolation, as reflected by g_C (1.31), g_S (1.4) and g_{lin} (1.68). As illustrated in Figure 3 (left), the total squared error is smaller for E_{lin} than for C at any x, and it is smaller than that for F_* almost everywhere, with exception of narrow intervals near x = 0 and x = 1. In addition, E_{lin} loses to S around x = 0.5, but wins in wide intervals near x = 0 and x = 1.

More interesting results follow if linear interpolation is made according to eq. (17). Now the interpolation target is x, i.e. ecdf-values are selected as an independent variable *e*. In this case the implicitly defined ecdf *e*(x, X) is given by:

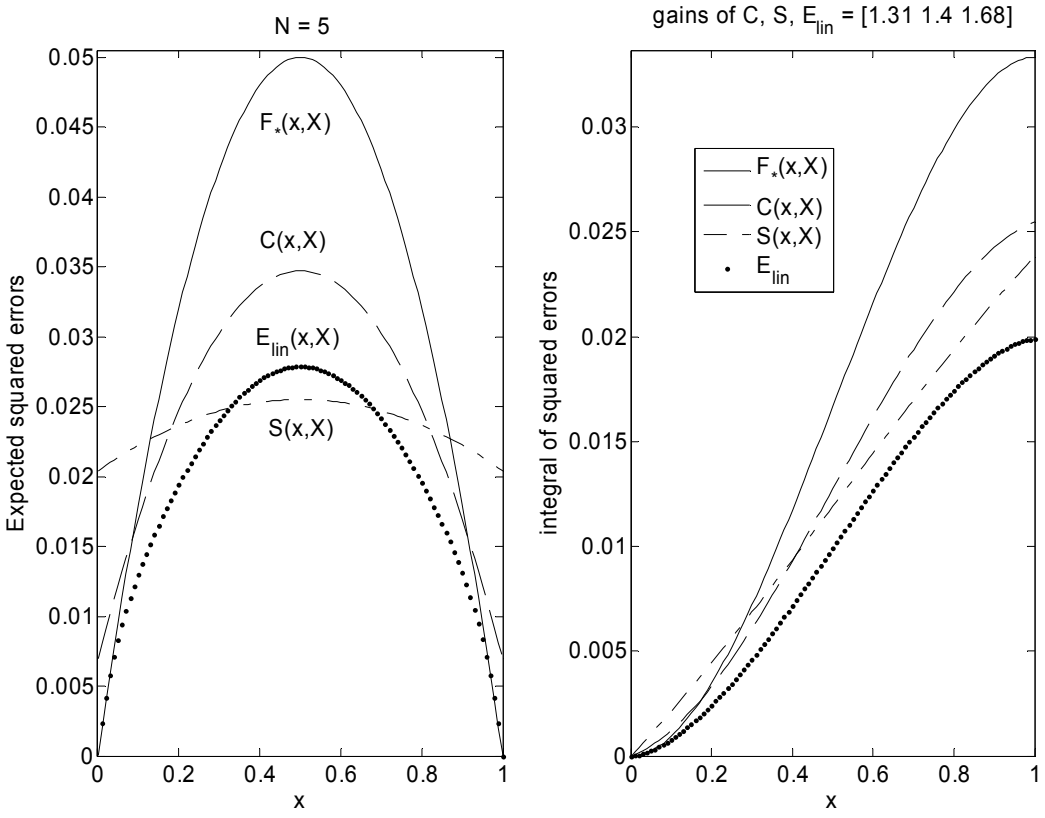
$$x(e, X) = \mathbf{X}_{n-1}(1-\lambda)+ \mathbf{X}_n\lambda;\ n = 1:N+1. \quad (50)$$

Here, λ is the interpolation variable,

$$\lambda=(e-\mathbf{E}_{n-1})(N+1),\ 0 \le \lambda \le 1\ (\mathbf{E}_{n-1} \le e \le \mathbf{E}_n). \quad (51)$$

Note that an equation similar to (50) was used by Hyndman & Fan (1996), eq. (1), who applied linear interpolation for calculating distribution quantiles. Due to the linearity of eq. (50) with respect to random X-values, this equation represents an unbiased empirical estimation of parent U(0, 1), that is, $E[x(e, X)] = e$, which

Figure 3: Expected Squared Errors of Different Versions of ecdf for Samples from U(0, 1), N=5
Left: $E[\Delta^2]$; right: integrals of left curves in (0, x), which define at x = 1 the total expected errors.



immediately follows from E[**X**] = **E**. This is interesting, because it shows that $F_*(x, X)$ is not the only possible unbiased estimation of F(x). The squared error of x_{lin} defined by (50) is:

$$E[\Delta_{lin}^2] = E[(x(e,\mathbf{X})-e)^2]$$
$$= E[((\mathbf{X}_{n-1} - \mathbf{E}_{n-1})(1-\lambda) +(\mathbf{X}_n-\mathbf{E}_n)\lambda)^2]$$
$$= \mathbf{V}_{n-1}(1-\lambda)^2+\mathbf{V}_n\lambda^2+2c(n,n+1)\lambda(1-\lambda),\ n = 1{:}N+1. \quad (52)$$

Here c = cov(**X**), a covariance matrix of extended sample **X**, and **V**=[0 V 0] is the variance (12), extended by the values $V_0 = V_{N+1} = 0$. As can be seen from eq. (52), expected squared approximation error in every interval $\mathbf{E}_{n-1} \le e \le \mathbf{E}_n$ is given by parabola, connecting adjacent points ($\mathbf{E}_n$, $\mathbf{V}_n$). This is illustrated in Figure 4. The integral of (52) in (0, e) is now represented by piecewise cubic parabolas.

The gain of linear interpolation is now the same as in Figure 3, that is, the linear gain is invariant with respect to the interpolation target. The value of the linear gain for N > 1 is well approximated by g = 1 + 6/(2N-1), which means about 300%/N savings on sample size in comparison with F_*. This raises the question about how such gain correlates with the quality of predictions based on linear interpolation.

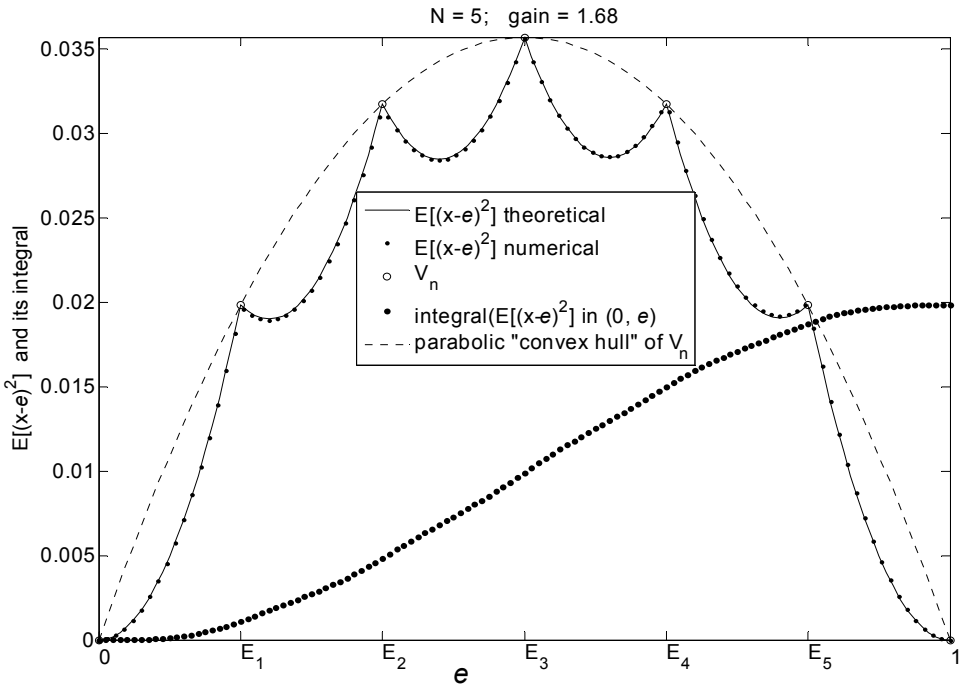
Eq. (8) can be directly applied to linear interpolation, which gives unbiased estimation and therefore eq. (8) should be valid. Given M = mean(X), eq. (8) suggests to represent x(*e*, X) as x(*e*, M(X)):

$$x = 2eM,\ e \le 0.5$$

and

$$x = 2(1-e)\,M + 2e-1,\ e > 0.5. \quad (53)$$

Figure 4: Expected Squared Error of the Linear Approximation (50), $E[(x-e)^2]$ and its Integral in (0, e)



Because E[M] = 0.5 (uniform data), (53) is indeed an unbiased estimation of x = e, the expected squared deviation of x from e is given by

$$\mathrm{var}(x) = 4e^2 V_M,\ e \le 0.5$$

and

$$\mathrm{var}(x) = 4(1-e)^2\, V_M,\ \ e > 0.5. \qquad (54)$$

Where

$$V_M = 1/(12N) \qquad (55)$$

V_M is the variance of M. Integrating (54) over e in (0, 1), the total mean squared deviation D_M is obtained as

$$D_M = 1/(36N). \qquad (56)$$

This result seems to be extraordinary, because it means a gain of ecdf (53) equal to 6, that is, 6 times shorter samples in comparison with F_* at the same approximation error, and this happens at any N value! Is it possible to get some practical advantages out of such a precise approximation?

One such possibility is suggested by distribution parameter fitting. Thus, unknown parameter(s) q can be found as argmin((mean(F(X, X, q))-0.5)^2).

This method works indeed, and it should be compared with others. However, fitting parameters is a special topic, which should be discussed separately.

Optimal ecdf S is constructed to minimize expected total error for continual applications. Discrete applications only need ecdf values at x = X, and then P(x, X) should be used. How is it possible to combine P(x, X) and S(x, X) into a universal ecdf, valid both for continual and discrete applications? This can be done by redefining S at x = X, e.g. by introducing function $\Phi(x, X) = S(x)$, if $x \ne X_n$, otherwise $\Phi(X_n, X) = E_n$, n = 1:N. Such switching between P(X, X) and S(x, X) can be expressed as a formal mixture of both functions, using Kronecker symbol δ_{xX}:

$$\Phi(x,X) = \delta_{xX}\, P(x,X) + (1- \delta_{xX})\, S(x,X),\ \delta_{xX} = 1,$$
$$\text{if any}(x==X), \text{otherwise } \delta_{xX} = 0. \qquad (57)$$

Function $\Phi(x, X)$ is discontinuous at x = X both from left and right, which is physically and functionally more reasonable, than in the case of $F_*(x, X)$, continuous from the right only.

Conclusion

The least error piecewise constant approximation of F(x) was presented. The starting point was that ecdf ascribes total probability of 1 to the sample, whereas any finite sample represents a set of measure zero. An optimal approach should ascribe zero measure associated with the sample. However, due to its convenience, a piecewise constant formalism has been selected. As a result, a part of total probability, equal to N/(N+2), is still associated with the sample. However, the aim was roughly achieved, because this measure is now smaller than 1, and this enabled a higher accuracy.

Optimal ecdf S(x, X) was built as a result of eliminating order bias of levels in ecdf $F_*(x, X)$, which is an unbiased estimation of F(x) for any fixed-in-advance x-value. Are ecdf versions C and S also unbiased? If it is forgotten for a moment that C and S are not designed for straightforward averaging over different samples, E[s(x, X, α)] could be calculated. As follows from (22), the s-family is biased at $\alpha > 0$, i.e. C and S are biased. This bias asymptotically disappears as $N \to \infty$. Is this bias important or not? What is more important for practical applications, improved accuracy of F(x) approximation, or formal bias which is in fact artificially created?

This bias has no practical meaning. Versions C and S use all available sample elements by definition, and the way this is done is not reducible to simple averaging. In fact, the bias is created by violation of the procedures behind C and S. The correct comparison is not reduced to an averaging over several samples. Instead, all available samples should be fused into one long sample before C or S functions are found. As eq. (8) shows, in the case of F_* the averaging over many samples gives the same result, as one combined sample. This enables formal ubiasedness, but the consequence thereof is increased approximation error.

A correct comparison of D_{F*}, D_C and D_S should always be done using the same sample or set of samples. If $N \to \infty$, then F_*, C and S all converge to the same F(x). The only difference is that D_S is the smallest of the three at any N. For this reason, if N is not very large, S(x, X) should always be preferred in practice as the best piece-wise constant approximation of F(x).

The smallest possible error of empirical estimation of F(x) is desirable, regardless of whether the error is due to the variance or due to inexact mathematical expectation. An optimal method should minimize the total error, and exactly this is done by $\Phi(x, X)$ both for discrete and continual applications. Physically, S has a smaller approximation error, because it takes into account additional information, contained in the order statistics F(X), whereas F_* neglects this information. As a result, ecdf F_* has order bias and an unnecessarily big approximation error.

The optimal ecdf $\Phi(x, X)$, presented here, is based on the most popular optimality criterion in statistics, i.e. least squared deviation. Final decision about its superiority depends on the quality of statistical predictions produced by different ecdf versions.

Acknowledgements

It is a pleasure to acknowledge Dr. L. Bogachev from the University of Leeds, UK, for his quite detailed discussions, Dr. J. Hilbe for several useful remarks, Dr. T. Lane from MathWorks, Inc. for discussions, Cleve Moler and the MathWorks team for the efficient software used in numerical simulations and Q. Beatty for improving my English.

References

Abramowitz, M., & Stegun, I. A. (1962). *Handbook of mathematical functions*. NY: Dover.

Brunk, H. D. (1962). On the range of the differences between hypothetical distribution function and Pyke's modified empirical distribution function. *The Annals of Mathematical Statistics*, 33(2), 525-532.

Cramér, H. (1971). *Mathematical methods of statistics*, (*12th Ed.*). Princeton, NJ: Princeton University Press.

Durbin, J. (1968). The probability that the sample distribution function lies between two parallel straight lines. *The Annals of Mathematical Statistics*, *39*(2), 398-411.

Durbin, J. (1973). Distribution theory for tests based on the sample distribution theory. *Regional Conference Series in Applied Mathematics*, *9*. UK: London School of Economics and Political Science, University of London.

Durbin, J., & Knott, M. (1972). Components of Cramér-von Mises statistics I. *Journal of the Royal Statistical Society*, *B*(*34*), 290-307.

Feller, W. (1971). *An introduction to probability theory and its applications*, (*2nd Ed.*). NY: John Wiley and Sons, Inc.

Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference*, (*3rd Ed.*). NY: M.Dekker.

Gumbel, E. J. (1939). La probabilité des Hypothèses. *Comptus Rendus de l'Academie des Sciences*, *209*. 645-647.

Hyndman, R. J., & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, *50*(4), 361-365.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C*, (*2nd Ed.*). Cambridge, MA: Cambridge University Press.

Pugachev, V. S. (1984). *Probability theory and mathematical statistics for engineers*. Elsevier Science

Ltd. In Russian: В.С. Пугачев. *Теория вероятностей и математическая статистика*.

Москва, Наука, Главная редакция физико-математической литературы, 1979, стр. 303.

Pyke, R. (1959). The supremum and infimum of the Poisson process. *Annals of Mathematical Statistics*, *30*, 568-576.

Weibull, W. (1939). The phenomenon of rupture in solids. *Ingeniörs Vetenskaps Akademien Handlinger*, *153*, 17.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 409-422

1538 – 9472/09/$95.00

Application of the Truncated Skew Laplace Probability Distribution in Maintenance System

Gokarna R. Aryal
Purdue University Calumet

Chris P. Tsokos
University of South Florida

A random variable X is said to have the skew-Laplace probability distribution if its pdf is given by $f(x) = 2g(x)G(\lambda x)$, where g (.) and G (.), respectively, denote the pdf and the cdf of the Laplace distribution. When the skew Laplace distribution is truncated on the left at 0 it is called it the truncated skew Laplace (TSL) distribution. This article provides a comparison of TSL distribution with two-parameter gamma model and the hypoexponential model, and an application of the subject model in maintenance system is studied.

Key words: Probability Distribution; Truncation; Simulation; Reliability, Renewal Process.

Introduction

Very few real world phenomena studied statistically are symmetrical in nature, thus, the symmetric models would not be useful for studying every phenomenon. The normal model is, at times, a poor description of observed phenomena. Skewed models, which exhibit varying degrees of asymmetry, are a necessary component of the modeler's tool kit. The term skew Laplace (SL) means a parametric class of probability distributions that extends the Laplace probability density function (pdf) by an additional shape parameter that regulates the degree of skewness, allowing for a continuous variation from Laplace to non-Laplace.

Gokarna Aryal is an Assistant Professor of Statistics in the Department of Mathematics, CS and Statistics at Purdue University Calumet. His research interest includes distribution theory, reliability analysis among others. Email: aryalg@calumet.purdue.edu. Chris Tsokos is a Distinguished University Professor in mathematics and Statistics at the University of South Florida. His research interests are in modeling Global Warming, analysis and modeling of cancer data, parametric, Bayesian and nonparametric reliability, and stochastic systems, among others. He is a fellow of both ASA and ISI. Email: profcpt@cas.usf.edu.

The skew Laplace distribution as a generalization of the Laplace law should be a natural choice in all practical situations in which some skewness is present. Several asymmetric forms of the skewed Laplace distribution have appeared in the literature with different formulations. Aryal *et al.* (2005b) studied extensively the mathematical properties of a skew Laplace distribution. This distribution was developed using the idea introduced by O'Hagan and studied by Azzalini (1985). A random variable X is said to have the skew symmetric distribution if its probability density function (pdf) is given by

$$f(x) = 2g(x)G(\lambda x) \qquad (1.1)$$

where, $-\infty < x < \infty$, $-\infty < \lambda < \infty$, $g(x)$ and $G(x)$ are the corresponding pdf and the cumulative distribution function (cdf) of the symmetric distributions.

The Laplace distribution has the pdf and cdf specified by

$$g(x) = \frac{1}{2\varphi}\exp\left(-\frac{|x|}{\varphi}\right) \qquad (1.2)$$

and

$$G(x)=\begin{cases}\frac{1}{2}\exp\left(\frac{x}{\varphi}\right) & if\ x\le 0,\\ 1-\frac{1}{2}\exp\left(-\frac{x}{\varphi}\right) & if\ x\ge 0\end{cases} \quad (1.3)$$

respectively, where $-\infty<x<\infty$ and $\varphi>0$. Hence, the pdf $f(x)$ and the cdf $F(x)$ of the skew Laplace random variable is given, respectively, by

$$f(x)=\begin{cases}\frac{1}{2\varphi}\exp\left\{-\frac{(1+|\lambda|)|x|}{\varphi}\right\}, if\ \lambda x\le 0,\\ \frac{1}{\varphi}\exp\left(-\frac{|x|}{\varphi}\right)\left\{1-\frac{1}{2}\exp\left(-\frac{\lambda x}{\varphi}\right)\right\}, if\ \lambda x\ge 0,\end{cases} \quad (1.4)$$

and

$$F(x)=\begin{cases}\frac{1}{2}+\frac{sign(\lambda)}{2}\left[\frac{1}{1+|\lambda|}\exp\left\{-\frac{(1+|\lambda|)|x|}{\varphi}\right\}-1\right], \\ \quad if\ \lambda x\le 0,\\ \frac{1}{2}+sign(\lambda)\left[\frac{1}{2}-\exp\left(-\frac{|x|}{\varphi}\right)\Phi(\lambda)\right], \\ \quad if\ \lambda x\ge 0.\end{cases} \quad (1.5)$$

where,

$$\Phi(\lambda)=1-\frac{1}{2(1+|\lambda|)}\exp\left(-\frac{\lambda x}{\varphi}\right).$$

Aryal et al. (2005a) proposed a reliability model that can be derived from the skew Laplace distribution on truncating it at 0 on the left. This is called the truncated skew Laplace (TSL) probability distribution. The cdf of this reliability model for $\lambda>0$ is given by

$$F^*(x)=1+\frac{\exp\left(-\frac{(1+\lambda)x}{\varphi}\right)-2(1+\lambda)\exp\left(-\frac{x}{\varphi}\right)}{(2\lambda+1)} \quad (1.6)$$

and the corresponding pdf is given by

$$f^*(x)=\begin{cases}\frac{(1+\lambda)}{\varphi(2\lambda+1)}\left\{2\exp\left(-\frac{x}{\varphi}\right)-\exp\left(-\frac{(1+\lambda)x}{\varphi}\right)\right\} if\ x\ge 0\\ 0 \qquad otherwise\end{cases} \quad (1.7)$$

It is immediate that the reliability function $R(t)$ and the hazard rate function $h(t)$ of a TSL random variable is given, respectively, by

$$R(t)=\frac{2(1+\lambda)\exp\left(-\frac{t}{\varphi}\right)-\exp\left(-\frac{(1+\lambda)t}{\varphi}\right)}{(2\lambda+1)} \quad (1.8)$$

and

$$h(t)=\frac{(1+\lambda)}{\varphi}\,\frac{\left\{2-\exp\left(\frac{-\lambda t}{\varphi}\right)\right\}}{\left\{2+2\lambda-\exp\left(\frac{-\lambda t}{\varphi}\right)\right\}}. \quad (1.9)$$

Also, note that the mean residual lifetime (MRLT) of a TSL random variable is given by

$$m(t)=\frac{\varphi}{(1+\lambda)}\left\{\frac{2(1+\lambda)^2-\exp\left(-\frac{\lambda t}{\varphi}\right)}{2(1+\lambda)-\exp\left(-\frac{\lambda t}{\varphi}\right)}\right\}. \quad (1.10)$$

This article provides a comparison of this reliability model with other competing models, namely, the two parameter gamma and hypoexponential distribution. We also study an application of the TSL probability model in preventive maintenance and cost optimization.

TSL vs. Gamma Distribution

A random variable X is said to have a gamma probability distribution with parameters α and β denoted by $G(\alpha,\beta)$ if it has a probability density function given by

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right),\ \alpha,\beta,x \geq 0, \tag{2.1}$$

where $\Gamma(.)$ denotes the gamma function. The parameters α and β are the shape and scale parameters, respectively. The reliability and hazard functions are not available in closed form unless α is an integer; however, they may be expressed in terms of the standard incomplete gamma function $\Gamma(a,z)$ defined by

$$\Gamma(a,z) = \int_0^z y^{a-1}\exp(-y)dy,\ a>0.$$

In terms of $\Gamma(a,z)$ the reliability function for random variable Gamma is given by

$$R(t;\alpha,\beta) = \frac{\Gamma(\alpha)-\Gamma(\alpha,t/\beta)}{\Gamma(\alpha)}. \tag{2.2}$$

If α is an integer, then the reliability function is given by

$$R(t;\alpha,\beta) = \sum_{k=0}^{\alpha-1}\frac{(t/\beta)^k\exp(-t/\beta)}{k!}. \tag{2.3}$$

The hazard rate function is given by

$$h(t;\alpha,\beta) = \frac{t^{\alpha-1}\exp(-t/\beta)}{\beta^{\alpha}\left[\Gamma(\alpha)-\Gamma(\alpha,t/\beta)\right]}, \tag{2.4}$$

for any $\alpha>0$, however, if α is an integer it becomes

$$h(t;\alpha,\beta) = \frac{t^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)\sum_{k=0}^{\alpha-1}(t/\beta)^k/k!}. \tag{2.5}$$

The shape parameter α is of special interest, since whether $\alpha-1$ is negative, zero or positive, corresponds to a decreasing failure rate (DFR), constant, or increasing failure rate (IFR), respectively.

It is clear that the gamma model has more flexibility than the TSL model as the former can be used even if the data has DFR. In fact, the standard exponential distribution is $TSL(0,1)$ as well as $Gamma(1,1)$. However, if in the gamma model $\alpha>1$, it has IFR which appears to be the same as that of the TSL model, but a careful study has shown a significance difference between these two models, this is the case for which real world data - where the TSL model gives a better fit than the competing gamma model - could be presented.

According to Pal *et al.* (2006) the failure times (in hours) of pressure vessels constructed of fiber/epoxy composite materials wrapped around metal lines subjected to a certain constant pressure, studied by Keating *et al.* (1990), can be described using $Gamma(1.45,300)$ model. The subject data was studied using TSL model. It was observed that TSL (5939.8, 575.5) fits the subject data better than the gamma distribution The Kolmogorov-Smirnov goodness of fit indicated that, the D-statistic for Gamma (1.45, 300) and TSL (5939.8, 575.5) distribution are $D_{Gamma}=0.2502$ *and* $D_{TSL}=0.200$ respectively. Since the smaller D-statistic, the better is the fit so it is concluded that the TSL model fits better than the gamma model.

Figure 1 displays the P-P plot of the fits of the pressure vessels data assuming the TSL and the gamma models. It is clear that the TSL pdf is a better fit than the gamma model. Thus, the TSL is recommended for the pressure vessel data. Table 1 gives the reliability estimates using TSL and gamma models. It is observed that there is a significant difference in these estimates.

TSL vs. Hypoexponential Probability Distribution

Observing the probability structure of the truncated skew Laplace pdf it is of interest to seek an existing probability distribution, which can be written as a difference of two exponential functions. Since the hypoexponential distribution has this characteristic the TSL pdf will be compared with the hypoexponential pdf. Many natural phenomena can be divided into

Figure 1: P-P Plots of Vessel Data Using TSL and Gamma Distribution

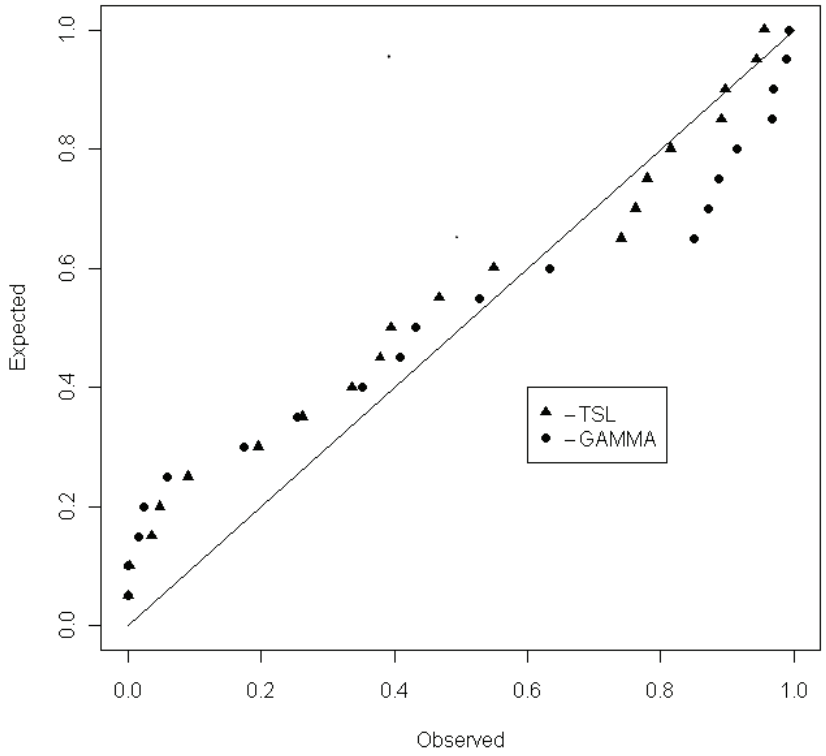


Table 1: Reliability Estimates of the Pressure Vessels Data

t	$\hat{R}_{TSL}(t)$	$\hat{R}_{GAMMA}(t)$	t	$\hat{R}_{TSL}(t)$	$\hat{R}_{GAMMA}(t)$
0.75	0.999	0.999	363	0.532	0.471
1.70	0.997	0.999	458	0.451	0.365
20.80	0.965	0.984	776	0.260	0.150
28.50	0.952	0.976	828	0.237	0.129
54.90	0.909	0.940	871	0.220	0.113
126.0	0.803	0.826	970	0.185	0.085
175.0	0.738	0.745	1278	0.108	0.034
236.0	0.664	0.647	1311	0.102	0.030
274.0	0.621	0.590	1661	0.056	0.010
290.0	0.604	0.567	1787	0.045	0.007

sequential phases. If the time the process spent in each phase is independent and exponentiallydistributed, then it can be shown that overall time is hypoexponentially distributed. It has been empirically observed that service times for input-output operations in a computer system often possess this distribution (see Trivedi, 1982) and will have n parameters one for each of its distinct phases. Interest then lies in a two-stage hypoexponential process, that is, if X is a random variable with parameters λ_1 and λ_2 $(\lambda_1 \neq \lambda_2)$ then its pdf is given by

$$f(x) = \frac{\lambda_1\lambda_2}{\lambda_2 - \lambda_1}\{\exp(-\lambda_1 x) - \exp(-\lambda_2 x)\},\ x > 0 \tag{3.1}$$

The notation $Hypo(\lambda_1, \lambda_2)$ denotes a hypoexponential random variable with parameters λ_1 and, λ_2 . The corresponding cdf is given by

$$F(x) = 1 - \frac{\lambda_2}{\lambda_2 - \lambda_1}\exp(-\lambda_1 x) + \frac{\lambda_1}{\lambda_2 - \lambda_1}\exp(-\lambda_2 x),\quad x \geq 0 \tag{3.2}$$

The reliability function $R(t)$ of a $Hypo(\lambda_1, \lambda_2)$ random variable is given by

$$R(t) = \frac{\lambda_2}{\lambda_2 - \lambda_1}\exp(-\lambda_1 t) - \frac{\lambda_1}{\lambda_2 - \lambda_1}\exp(-\lambda_2 t) \tag{3.3}$$

The hazard rate function $h(t)$ of a $Hypo(\lambda_1, \lambda_2)$ random variable is given by

$$h(t) = \frac{\lambda_1\lambda_2[\exp(-\lambda_1 t) - \exp(-\lambda_2 t)]}{\lambda_2\exp(-\lambda_1 t) - \lambda_1\exp(-\lambda_2 t)}. \tag{3.4}$$

It is clear that $h(t)$ is an increasing function of the parameter λ_2; it increases from 0 to $\min\{\lambda_1, \lambda_2\}$. Note that the mean residual life time (MRLT) at time t for $Hypo(\lambda_1, \lambda_2)$ is given by

$$m_{Hypo}(t) = \frac{1}{\lambda_1\lambda_2}\frac{\lambda_2^2\exp(-\lambda_1 t) - \lambda_1^2\exp(-\lambda_2 t)}{[\lambda_2\exp(-\lambda_1 t) - \lambda_1\exp(-\lambda_2 t]}. \tag{3.5}$$

To compare the TSL and hypoexponential pdf in terms of reliability and mean residual life times, random samples of size 50, 100 and 500 are generated from a hypoexponential pdf with parameters λ_1=1 and $\lambda_2 = 2, 5, 10 \& 20$ for each sample size. Numerical iterative procedure, Newton-Raphson algorithm, is used to estimate the maximum likelihood estimates of λ_1 & λ_2. To compare these results, the parameters φ and λ of a TSL distribution are estimated (See Table 2). In addition the mean residual life times were computed for both the models at $t = t_{n/2}$.

In Table 2, M_{TSL} and M_{HYPO} denote the MRLT of TSL and hypoexponential models respectively. Table 2 shows that if the sample size is large and the difference between the two parameters λ_1 and λ_2 is large both the TSL and hypoexponential model will produce the same result. However, for a small sample size and a small difference between λ_1 and λ_2 a significant difference is observed between the two models. Figures 2-5 illustrate the plotted reliability graphs and provide the support for these findings.

TSL Distribution and Preventive Maintenance

In many situations, failure of a system or unit during actual operation can be very costly or in some cases dangerous if the system fails, thus, it may be better to repair or replace before it fails. However, it is not typically feasible to make frequent replacements of a system. Thus, developing a replacement policy that balances the cost of failures against the cost of planned replacement or maintenance is necessary. Suppose a unit that is to operate over a time 0 to time t, [0, t], is replaced upon failure (with failure probability distribution F). Assume

Table 2: Mean Residual Lifetimes (MRLT) of TSL and Hypoexponential Models Computed by Using (1.10) and (3.5) for Different Sample Sizes

n	λ_1	λ_2	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\varphi}$	$\hat{\lambda}$	M_{TSL}	M_{HYPO}
50	1	2	0.934	2.325	2.745	1.349	1.362	1.129
50	1	5	0.975	5.133	2.779	1.097	1.108	1.029
50	1	10	0.979	12.223	1.042	6.968	1.042	1.021
50	1	20	0.940	26.742	1.069	15.349	1.069	1.063
100	1	2	0.876	2.565	1.376	2.403	1.391	1.184
100	1	5	0.903	6.835	1.178	6.216	1.179	1.108
100	1	10	0.950	9.838	1.098	8.439	1.099	1.052
100	1	20	1.029	26.322	0.892	0.242	0.982	0.971
500	1	2	0.915	2.576	1.339	3.076	1.348	1.132
500	1	5	0.961	6.489	1.088	3.453	1.093	1.042
500	1	10	0.881	10.224	1.174	8.355	1.173	1.135
500	1	20	1.016	27.411	0.988	14.044	0.988	0.983

Figure 2: Reliability of TSL and Hypoexponential Distributions for n=50, λ_1=1, λ_2=2

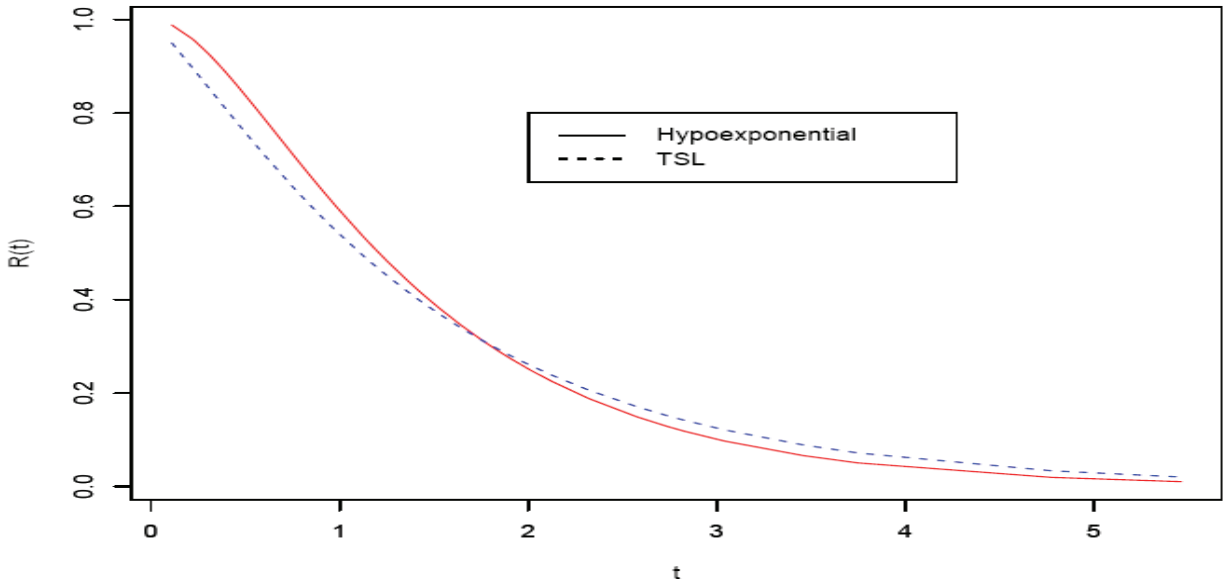


Figure 3: Reliability of TSL and Hypoexponential Distributions for n=50, λ_1=1, λ_2=5

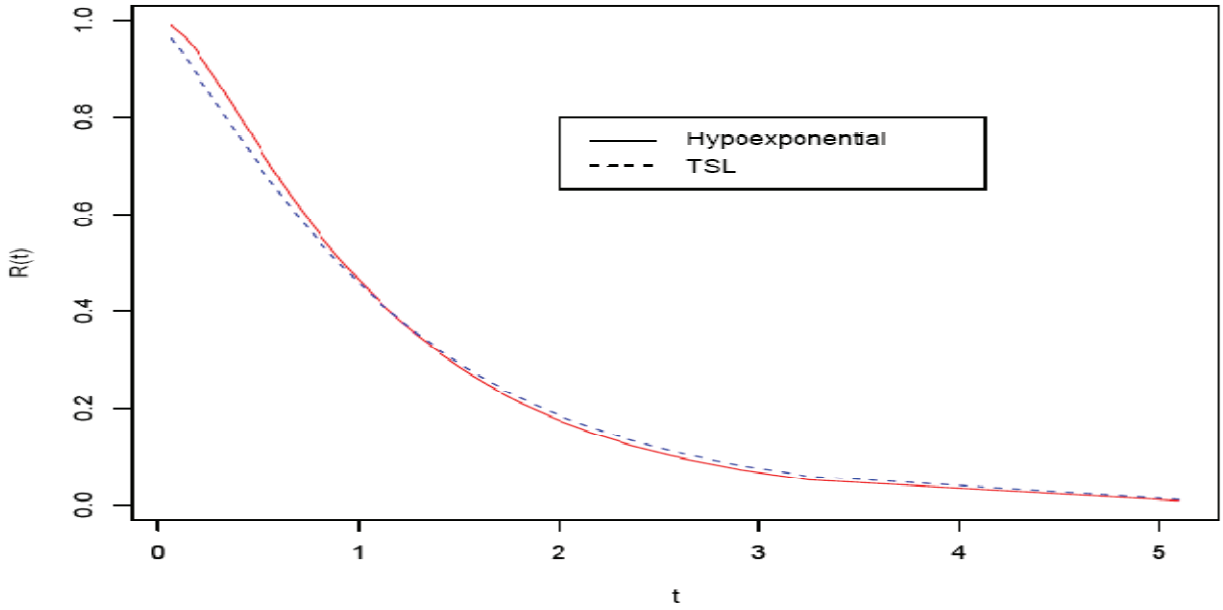


Figure 4: Reliability of TSL and Hypoexponential Distributions for n=50, λ_1=1, λ_2=10

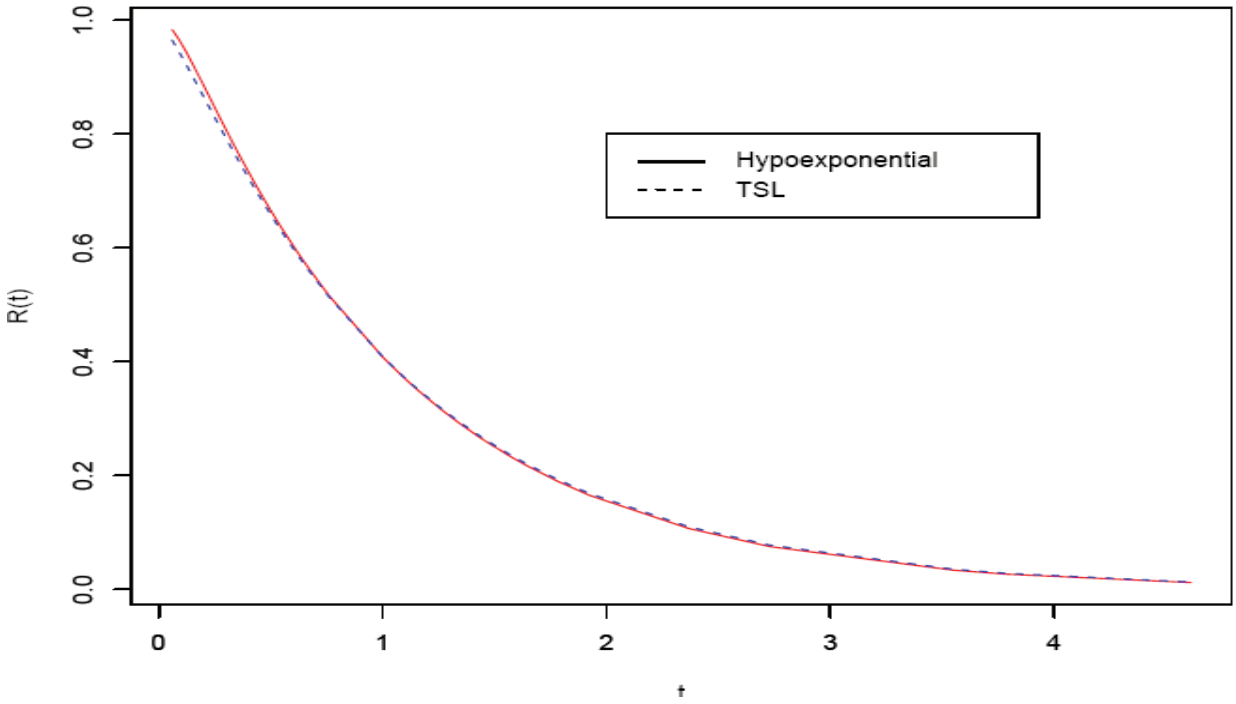
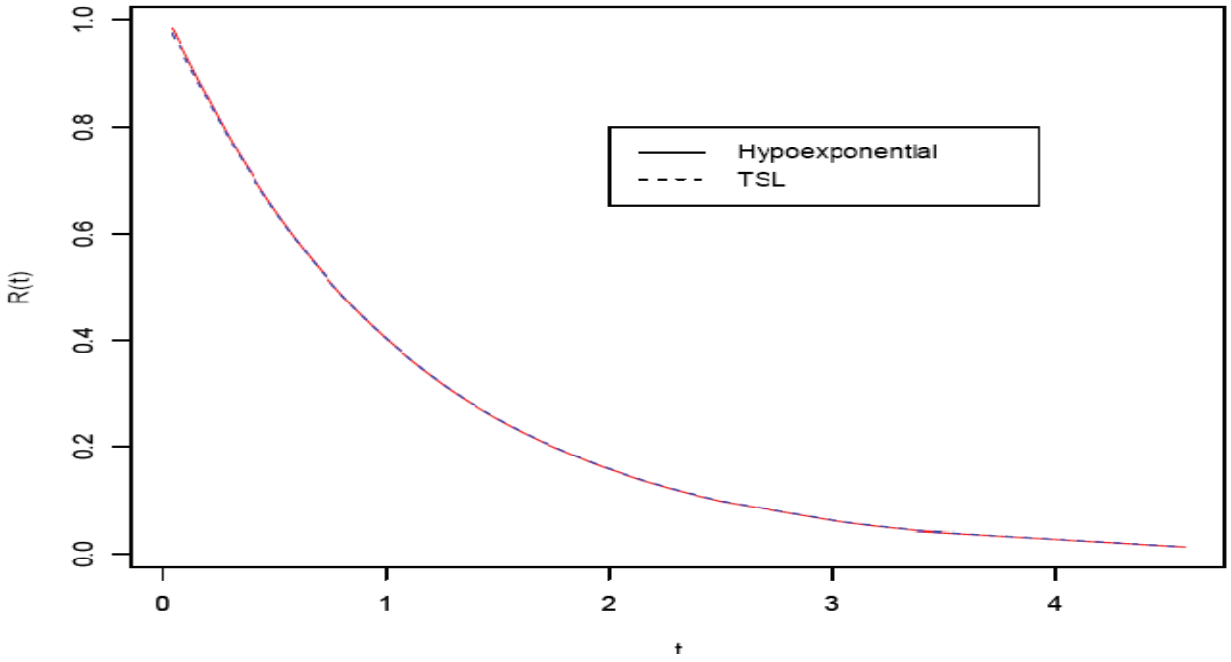


Figure 5: Reliability of TSL and Hypoexponential Distributions for n=50, λ_1=1, λ_2=20



that the failures are easily detected and instantly replaced and that cost c_1 includes the cost resulting from planned replacement and cost c_2 that includes all costs invested resulting from failure then the expected cost during the period [0, t] is given by

$$C(t) = c_1 E(N_1(t)) + c_2 E(N_2(t)), \quad (4.1)$$

where, $E(N_1(t))$ and $E(N_2(t))$ denote the expected number of planned replacement and the expected number of failures respectively. The goal is to determine the policy minimizing $C(t)$ for a finite time span or minimizing $\lim_{t\to\infty} \frac{C(t)}{t}$ for an infinite time span. Because the TSL probability distribution has an increasing failure rate it is expected that this model would be useful in a maintenance system.

Age Replacement Policy and TSL Probability Distribution

Consider the so-called age replacement policy; in this policy an item is always replaced exactly at the time of failure, or at t^* time after installation, whichever occurs first. This age replacement policy for infinite time spans seems to have received the most attention in the literature. Morese (1958) showed how to determine the replacement interval minimizing cost per unit time, while Barlow and Proschen (1962) proved that if the failure distribution, F, is continuous then a minimum-cost age replacement exists for any infinite time span.

In this article, the goal is to determine the optimal time t^* at which preventive replacement should be performed. The model should determine the time t^* that minimizes the total expected cost of preventive and failure maintenance per unit time. The total cost per cycle consists of the cost of preventive maintenance in addition to the cost of failure

maintenance. Hence,

$$EC(t^*) = c_1\left(R(t^*)\right) + c_2\left(1 - R(t^*)\right) \quad (4.1.1)$$

where c_1 and c_2 denote the cost of preventive maintenance and failure maintenance respectively, and $R(t^*)$ is the probability that the equipment survives until age t^*. The expected cycle length consists of the length of a preventive cycle plus the expected length of a failure cycle. Thus, we have

$$\text{Expected Cycle Length} = t^* R(t^*) + M(t^*)\left(1 - R(t^*)\right) \quad (4.1.2)$$

where

$$M(t^*)\left(1 - R(t^*)\right) = \int_{-\infty}^{t^*} tf(t)dt$$

is the mean of the truncated distribution at time t^*. Hence, the expected cost per unit time is equal to:

$$\frac{c_1 R(t^*) + c_2\left[1 - R(t^*)\right]}{t^* R(t^*) + M(t^*)\left[1 - R(t^*)\right]} \quad (4.1.3)$$

Assume that a system has a time to failure distribution of the truncated skew Laplace pdf; the goal is to compute the optimal time t^* of preventive replacement. Because the reliability function of a TSL random variable is given by

$$R(t^*) = \frac{2(1+\lambda)\exp\left(-\frac{t^*}{\varphi}\right) - \exp\left(-\frac{(1+\lambda)t^*}{\varphi}\right)}{(2\lambda+1)}$$

and

$$M(t^*) = \frac{1}{1 - R(t^*)}\int_0^{t^*} tf(t)dt$$

thus,

$$\begin{aligned}\int_0^{t^*} tf(t)dt = {} & \frac{2(1+\lambda)\varphi}{2\lambda+1}\left[1 - \exp(-t^*/\varphi)\right] \\ & - \frac{2(1+\lambda)}{2\lambda+1} t^* \exp(-t^*/\varphi) \\ & + \frac{t^*}{(2\lambda+1)}\exp\left(-(1+\lambda)t^*/\varphi\right) \\ & - \frac{\varphi}{(2\lambda+1)(1+\lambda)t^*/\varphi}\left[1 - \exp\left(-(1+\lambda)\frac{t^*}{\varphi}\right)\right]\end{aligned}$$

Substituting and simplifying the expressions, the expected cost per unit time (ECU) is given by:

$$ECU(t^*) = \frac{(1+\lambda)}{\varphi}\frac{\begin{Bmatrix} 2(c_2 - c_1)(1+\lambda)\exp(-t^*/\varphi) \\ -(c_2 - c_1)\exp\left(-(1+\lambda)t^*/\varphi\right) \\ -c_2(2\lambda+1)\end{Bmatrix}}{\begin{Bmatrix} 2(1+\lambda)^2\exp\left(-t^*/\varphi\right) \\ -(1+4\lambda+2\lambda^2) \\ -\exp\left(-(1+\lambda)t^*/\varphi\right)\end{Bmatrix}} \quad (4.1.4)$$

Methodology

In order to minimize a function $g(t)$ subject to $a \le t \le b$ the Golden Section Method, which employs the following steps to calculate the optimum value may be used.

Step 1:

Select an allowable final tolerance level δ and assume the initial interval where the minimum lies is $[a_1, b_1] = [a, b]$ and let

$$\lambda_1 = a_1 + (1-\alpha)(b_1 - a_1)$$
$$\mu_1 = a_1 + \alpha(b_1 - a_1)$$

Take $\alpha = 0.618$, which is a positive root of $c^2 + c - 1 = 0$, evaluate $g(\lambda_1)$ and $g(\mu_1)$, and let $k = 1$. Go to Step 2.

Step 2:
If $b_k - a_k \le \delta$, stop because the optimal solution is $t^* = (a_k + b_k)/2$, otherwise, if $g(\lambda_k) > g(\mu_k)$ go to Step 3; or if $g(\lambda_k) \le g(\mu_k)$, go to Step 4.

Step 3:
Let $a_{k+1} = a_k$, $b_{k+1} = b_k$, $\lambda_{k+1} = \mu_k$ and $\mu_{k+1} = a_{k+1} + \alpha(b_{k+1} - a_{k+1})$. Evaluate $g(\mu_{k+1})$ and go to Step 5.

Step 4:
Let $a_{k+1} = a_k$, $b_{k+1} = \mu_k$, $\mu_{k+1} = \lambda_k$ and $\lambda_{k+1} = a_{k+1} + (1-\alpha)(b_{k+1} - a_{k+1})$. Evaluate $g(\lambda_{k+1})$ and go to Step 5.

Step 5:
Replace k by $k+1$ and go to Step 1.

Example
To implement this method to find the time t^* subject to the condition that $c_1 = 1$ and $c_2 = 10$ proceed as follows:

Iteration 1:
Consider $[a_1, b_1] = [0, 10]$, where $\alpha = 0.618$ so that $1 - \alpha = 0.382$.

$$\lambda_1 = a_1 + (1-\alpha)(b_1 - a_1) = 3.82$$
$$\mu_1 = a_1 + \alpha(b_1 - a_1) = 6.18,$$
$$ECU(\lambda_1) = 8.561, \text{ and } ECU(\mu_1) = 8.570.$$

Because $ECU(\lambda_1) < ECU(\mu_1)$ the next interval where the optimal solution lies is $[0, 6.18]$.

Iteration 2:
$[a_2, b_2] = [0, 6.18], \lambda_2 = 2.36$ *and* $\mu_2 = 3.82$.
$ECU(\lambda_2) = 8.533$ *and* $ECU(\mu_2) = 8.561$.

Because $ECU(\lambda_2) < ECU(\mu_2)$ the next interval where the optimal solution lies is $[0, 3.82]$.

Iteration 3:
$[a_3, b_3] = [0, 3.82], \lambda_3 = 1.459$ *and* $\mu_3 = 2.36$
$ECU(\lambda_3) = 8.516$ *and* $ECU(\mu_3) = 8.533$.

Because $ECU(\lambda_3) < ECU(\mu_3)$ the next interval where the optimal solution lies is $[0, 2.36]$.

Iteration 4:
$[a_4, b_4] = [0, 2.36], \lambda_4 = 0.901$ *and* $\mu_4 = 1.459$
$ECU(\lambda_4) = 8.613$ *and* $ECU(\mu_4) = 8.516$.

Because $ECU(\lambda_4) > ECU(\mu_4)$ the next interval where the optimal solution lies is $[0.901, 2.36]$.

Iteration 5:
$[a_5, b_5] = [0.901, 2.36], \lambda_5 = 1.459$ *and* $\mu_5 = 1.803$
$ECU(\lambda_5) = 8.516$ *and* $ECU(\mu_5) = 8.517$.

Because $ECU(\lambda_5) < ECU(\mu_5)$ the next interval where the optimal solution lies is $[0.901, 1.803]$.

Iteration 6:
$[a_6, b_6] = [0.901, 1.803], \lambda_6 = 1.246$ *and* $\mu_6 = 1.459$
$ECU(\lambda_6) = 8.528$ *and* $ECU(\mu_6) = 8.516$.

Because $ECU(\lambda_6) > ECU(\mu_6)$ the next interval where the optimal solution lies is $[1.246, 1.803]$.

Iteration 7:
$[a_7, b_7] = [1.246, 1.803], \lambda_7 = 1.459$ *and* $\mu_7 = 1.590$
$ECU(\lambda_7) = 8.516$ *and* $ECU(\mu_7) = 8.514$.

Because $ECU(\lambda_7) > ECU(\mu_7)$ the next interval where the optimal solution lies is $[1.459,\ 1.803]$.

If the δ level is fixed at 0.5, it can be concluded that the optimum value lies in the interval $[1.459,\ 1.803]$ and is given by 1.631. This numerical example was performed assuming that the failure data follows the $TSL(1,1)$ model and it has been observed that to optimize the cost, maintenance should be scheduled at 1.631 units of time.

Block Replacement Policy and TSL Probability Distribution

Consider the case of the Block-Replacement Policy, or the constant interval policy. In this policy preventive maintenance is performed on the system after it has been operating a total of t^* units of time, regardless of the number of intervening failures. In the case where the system has failed prior to the time t^*, minimal repairs are be performed. Assume that the minimal repair will not change the failure rate of the system and that preventive maintenance renews the system to its original new state. Thus, the goal is to find the time t^* that minimizes the expected repair and preventive maintenance costs. The total expected cost per unit time for preventive replacement at time t^*, denoted by ECU (t^*) is given by

$$ECU(t^*) = \frac{Total\ \text{exp}ected\ \text{cos}t\ in\ the\ \text{int}erval(0,t^*)}{Length\ of\ the\ \text{int}erval} \tag{4.2.1}$$

The total expected cost in the interval $(0,\ t^*)$ equals the cost of preventative maintenance plus the cost of failure maintenance, which is given by $c_1 + c_2 M(t^*)$, where M (t*) is the expected number of failures in the interval (0, t*). Thus,

$$ECU(t^*) = \frac{c_1 + c_2 M(t^*)}{t^*}. \tag{4.2.2}$$

It is known that the expected number of failures in the interval (0, t*) is the integral of the failure rate function, that is

$$M(t^*) = E\left(N(t^*)\right) = H(t^*) = \int_0^{t^*} h(t)dt.$$

Therefore, if the failure of the system follows the TSL distribution it may be observed that

$$M(t^*) = \int_0^{t^*} h(t)dt = \frac{(1+\lambda)t^*}{\varphi} - \log\left((2+2\lambda)\exp(\lambda t^*/\varphi) - 1\right) + \log(2\lambda+1)$$

Therefore,

$$ECU(t^*) = \frac{c_1 + c_2\left[\frac{(1+\lambda)t^*}{\varphi} - \log\left((2+2\lambda)\exp(\lambda t^*/\varphi) - 1\right) + \log(2\lambda+1)\right]}{t^*}. \tag{4.2.3}$$

Example

To minimize the total expected cost subject to the conditions $c_1 = 1$ and $c_2 = 10$, the Golden Section Method (as described above) is used to obtain the value of t^*

Iteration 1:

$[a_1, b_1] = [0,10], \alpha = 0.618$, $1-\alpha = 0.382$,

$$\lambda_1 = a_1 + (1-\alpha)(b_1 - a_1) = 3.82,$$
$$\mu_1 = a_1 + \alpha(b_1 - a_1) = 6.18,$$

$ECU(\lambda_1) = 9.523$ and $ECU(\mu_1) = 9.697.$

Because $ECU(\lambda_1) < ECU(\mu_1)$ the next interval where the optimal solution lies is $[0,\ 6.18]$.

Iteration 2:
$[a_2, b_2] = [0, 6.18]$, $\lambda_2 = 2.36$, $\mu_2 = 3.82$, $ECU(\lambda_2) = 9.30$, and $ECU(\mu_2) = 9.523$.

Because $ECU(\lambda_2) < ECU(\mu_2)$ the next interval where the optimal solution lies is $[0,\ 3.82]$.

Iteration 3:
$[a_3, b_3] = [0, 3.82]$, $\lambda_3 = 1.459$, $\mu_3 = 2.36$, $ECU(\lambda_3) = 9.124$ and $ECU(\mu_3) = 9.30$.

Because $ECU(\lambda_3) < ECU(\mu_3)$ the next interval where the optimal solution lies is $[0,\ 2.36]$.

Iteration 4:
$[a_4, b_4] = [0, 2.36]$, $\lambda_4 = 0.901$, $\mu_4 = 1.459$, $ECU(\lambda_4) = 9.102$ and $ECU(\mu_4) = 9.124$.

Because $ECU(\lambda_4) < ECU(\mu_4)$ the next interval where the optimal solution lies is $[0,\ 1.459]$.

Iteration 5:
$[a_5, b_5] = [0, 1.459]$, $\lambda_5 = 0.557$, $\mu_5 = 0.901$, $ECU(\lambda_5) = 9.405$ and $ECU(\mu_5) = 9.124$.

Because $ECU(\lambda_5) > ECU(\mu_5)$ the next interval where the optimal solution lies is $[0.557,\ 1.459]$.

Iteration 6:
$[a_6, b_6] = [0.557, 1.459]$, $\lambda_6 = 0.9015$, $\mu_6 = 1.114$, $ECU(\lambda_6) = 9.102$, and $ECU(\mu_6) = 9.08$.

Because $ECU(\lambda_6) > ECU(\mu_6)$ the next interval where the optimal solution lies is $[0.901,\ 1.459]$.

Iteration 7:
$[a_7, b_7] = [0.901, 1.459]$, $\lambda_7 = 1.114$, $\mu_7 = 1.245$, $ECU(\lambda_7) = 9.08$, and $ECU(\mu_7) = 9.09$.

Because $ECU(\lambda_7) < ECU(\mu_7)$ the next interval where the optimal solution lies is $[0.901,\ 1.245]$.

If the δ level is fixed at 0.5 it can be concluded that the optimum value lies in the interval $[0.901,\ 1.245]$ and it is given by 1.07. As in the case of age replacement in this numerical example it was assumed that the failure data follows $TSL(1,1)$ model. Observe that, in order to optimize the cost, maintenance must be scheduled at every 1.07 units of time.

Maintenance Over a Finite Time Span

The problem concerning the preventive maintenance over a finite time span is of great importance in industry. It can be viewed in two different perspectives based on whether the total number of replacements (failure + planned) times are known or unknown. The first case is straightforward and has been addressed in the literature for a long time. Barlow *et al.* (1962) derived the expression for this case. Let T^* represent the total time span, meaning minimization of the cost due to forced replacement or planned replacement until $T = T^*$. Let $C_n(T^*, T)$ represent the expected cost in the time span 0 to T^*, $[0, T^*]$, considering only the first n replacements and following a policy of replacement at interval T. It is clear that considering the case when $T^* \le T$ is equivalent to zero planned replacements, that

$$C_1(T^*, T) = \begin{cases} c_2 F(T^*), \text{ if } T^* \le T, \\ c_2 F(T) + c_1\big(1 - F(T)\big), \text{ if } T^* \ge T \end{cases} \tag{5.1}$$

Thus, for $n = 1, 2, 3, \ldots$, $C_{n+1}(T^*, T) =$

$$\begin{cases} \int_0^{T^*} \left[c_2 + C_n(T^* - y, T)\right] dF(y) \text{ if } T^* \le T \\ \int_0^{T} \left[c_2 + C_n(T^* - y, T)\right] dF(y) \\ + \left[c_1 + C_n(T^* - T, T)\right] \times \left[1 - F(T)\right], \text{ otherwise} \end{cases} \quad (5.2)$$

A statistical model may now be developed that can be used to predict the total cost of maintenance before an item is actually used. Let T equal the predetermined replacement time, and assume that an item is always replaced exactly at the time of failure T^* or T hours after its installation, whichever occurs first. Let τ denotes the failure time then we have two cases to consider,

Case 1: $T < T^*$

In this case the preventative maintenance (PM) interval is less than the finite planning horizon. For this case if the component fails after time, say, $\tau\,(for\ \tau < T)$ then the cost due to failure is incurred and the planning horizon is reduced to $[T^* - \tau]$. But if the component works till the preventive replacement time T then the cost due to preventive maintenance is incurred and the planning horizon is reduced to $[T^*-T]$.

The total cost incurred in these two cases is

$$C(T^*, T) = \int_0^{T} \left[c_2 + C(T^* - \tau, T)\right] f(\tau) d\tau + \left[1 - F(T)\right] \times \left[c_1 + C(T^* - T, T)\right] \quad (5.3)$$

where c_1 is the cost for preventive maintenance and $c_2 (> c_1)$ is the cost for failure maintenance.

Case 2: $T^* < T$

In this case the PM interval is greater than the planning horizon so there is no preventive maintenance but there is a chance of failure maintenance. Hence the total cost incurred will be

$$C(T^*, T) = \int_0^{T^*} \left[c_2 + C(T^* - \boldsymbol{\tau}, T)\right] f(\boldsymbol{\tau}) d\boldsymbol{\tau} \quad (5.4)$$

The interest here lies in finding the preventative maintenance time T that minimizes the cost of the system. Consider a numerical example to determine whether the minimum exists if the failure model is assumed to be TSL *(1, 1)*. A random sample of size 100 was generated from *TSL (1, 1)* and a time T to perform preventive maintenance was fixed. The preventive maintenance cost $c_1 = 1$ was set along with the failure replacement cost $c_2 = 1$, 2 and 10. The process was repeated several times and the total cost for first 40 failures was computed. All necessary calculations were performed using the statistical language R. In the table 3, C_i, for $i = 1, 2, \& 10$ represents the total cost due to preventive maintenance cost $c_1 = 1$ and the failure replacement cost $c_2 = i,\ i = 1, 2\ \&10$. It can be observed from table 3 that the minimum C_i exists around at T = 1.1 units of time. A preliminary study of the application of the TSL distribution in such environment can be found in Aryal, et al. (2008).

Table 3: Expected Costs at Different Maintenance Times

T	C_{10}	C_2	C_1
1.00	340.55	88.95	57.50
1.01	347.25	89.65	57.45
1.02	336.95	87.75	56.60
1.03	342.95	88.15	56.30
1.04	339.15	87.15	55.65
1.05	341.25	87.25	55.50
1.06	334.40	86.40	55.40
1.07	343.75	87.35	55.30
1.08	332.15	84.95	54.05
1.09	338.55	85.81	54.22
1.10	318.48	82.67	53.19
1.11	327.68	84.04	52.59
1.12	344.76	86.48	54.19
1.13	333.70	84.50	53.35
1.14	340.40	85.20	53.30
1.15	338.86	84.68	53.90
1.16	331.28	82.90	53.86
1.17	338.27	84.09	54.31
1.18	335.24	83.05	53.52
1.19	341.90	84.00	54.76
1.20	363.90	87.50	56.95

Conclusion

This study presented a comparison of the truncated skew Laplace probability distribution with the two parameter gamma probability distribution and hypoexponential probability distribution. A detailed procedure was also provided to apply the truncated skew Laplace probability distribution in the maintenance system.

Acknowledgements

The authors would like to express their sincere gratitude to the late Professor A. N. V. Rao for his useful suggestions and encouragements.

References

Aryal, G., & Rao, A. N. V. (2005a). *Reliability model using truncated skew Laplace distribution.* Nonlinear Analysis, 63, 639-646.

Aryal, G., & Nadarajah, S. (2005b). *On the skew Laplace distribution,* Journal of Information and Optimization Sciences, 26, 205-217.

Aryal, G., & Tsokos, C. P. (2008). *Theory and Applications of the Truncated Skew Laplace Distribution*, proceedings of Dynamic Systems and Applications, 5, 45-52.

Azzalini, A. (1985) *A class of distributions that includes the normal ones,* Scand. J. Statistics, 12, 171-178.

Barlow R. E., & Proschen, F. (1962). *Studies in applied probability and management science*. Stanford , CA: Stanford University Press.

Ihaka, R., & Gentleman, R. (1996). *R*: *A language for data analysis and graphics* Journal of Computational and Graphical Statistics, 5, 299-314.

Keating, J. P., Glaser, R. E., &Ketchum, N. S. (1990). *Testing hypothesis about the shape parameter of a Gamma distribution.* Technometrics, 32, 67-82.

Pal, N., Jin, C., & Lim, W. K. (2006). *Handbook of exponential and related distributions for engineers and scientists*. NY: Chapman& Hall/CRC.

Trivedi, K. S. (1982). *Probability and Statistics with reliability, queuing and computer science applications*. NY: Prentice-Hall, Inc.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 423-447

1538 – 9472/09/$95.00

On Some Discrete Distributions and their Applications with Real Life Data

Shipra Banik
Independent University, Bangladesh

B. M. Golam Kibria
Florida International University

This article reviews some useful discrete models and compares their performance in terms of the high frequency of zeroes, which is observed in many discrete data (e.g., motor crash, earthquake, strike data, etc.). A simulation study is conducted to determine how commonly used discrete models (such as the binomial, Poisson, negative binomial, zero-inflated and zero-truncated models) behave if excess zeroes are present in the data. Results indicate that the negative binomial model and the ZIP model are better able to capture the effect of excess zeroes. Some real-life environmental data are used to illustrate the performance of the proposed models.

Key words: Binomial Distribution; Poisson distribution; Negative Binomial; ZIP; ZINB.

Introduction

Statistical discrete processes – for example, the number of accidents per driver, the number of insects per leaf in an orchard, the number of thunderstorms per year, the number of earthquakes per year, the number of patients visit emergency room in a certain hospital per day - often occur in real life. To approximate (or fit) a process, statistical probabilistic distributions are often used. Thus, fitting a process has been drawn considerable attention in the literature of many fields, for example, engineering (Lord, et al., 2005), ecology (Warton, 2005), biological science (Lloyd-Smith, 2007; Bliss & Fisher, 1953), epidemiology (Bohning, 1998), entomology (Taylor, 1961), zoology (Fisher, 1941), bacteriology (Neyman, 1939).

A broad range of probability models are commonly used in applied literature to fit discrete processes. These include: binomial model, Poisson model, negative binomial model, zero-inflated models and zero-truncated models. Binomial distribution models represent the total number of successes in a fixed number of repeated trials when only two outcomes are possible on each trial. Poisson distributions approximate rare-event processes (e.g., accident occurrences, failures in manufacturing or processing, etc.). An important restriction of the Poisson distribution is that its mean and variance are equal.

In reality, discrete processes often exhibit a large variance and a small mean and thus, display over-dispersion with a variance-to-mean value greater than 1 (Bliss & Fisher, 1953; Warton, 2005; Ross & Preece, 1985; White & Bennetts, 1996). Therefore, in real life, the Poisson assumption is often violated. A negative binomial distribution may be used for modeling purposes because it uses an additional parameter to describe the variance of a variable. Hence, the negative binomial distribution is considered as the first alternative to the Poisson distribution when the process is over-dispersed.

However, in many situations (e.g., road crash data), the chance of observing zero is

Shipra Banik is an Assistant Professor in the School of Engineering and Computer Science, Independent University, Bangladesh. Email: banik@secs.iub.edu.bd. B. M. Golam Kibria is an Associate Professor in the Department of Mathematics and Statistics at the Florida International University. He is the overseas managing editor of the *Journal of Statistical Research*, coordinating editor for the *Journal of Probability and Statistical Science*. He is an elected fellow of the *Royal Statistical Society* and the *International Statistical Institute.* Email: kibriag@fiu.edu.

greater than expected. Reasons may include failing to observe an event during the observational period and an inability to ever experience an event. Some researchers (Warton, 2005; Shankar, et al., 2003; Kibria, 2006) have applied zero-inflated models to model this type of process (known as a dual-states process: one zero-count state and one other normal-count state). These models generally capture apparent excess zeroes that commonly arise in some discrete processes, such as road crash data, and improve statistical fit when compared to the Poisson and the negative binomial model. The reason is that data obtained from a dual-state process often suffer from over-dispersion because the number of zeroes is inflated by the zero-count state. A zero-inflated model (introduced by Rider, 1961) is defined by

$$P(X=k)=\begin{cases}1-\theta & \text{if } X=0\\ \theta P(X;\mu_i) & \text{if } X>0\end{cases}$$

where θ is the proportion of non-zero values of X and $P(X;\mu_i)$ is a zero-truncated probability model fitted to normal-count states. To address phenomena with zero-inflated counting processes, the zero-inflated Poisson (ZIP) model and the zero-inflated negative binomial (ZINB) model have been developed. A ZIP model is a mix of a distribution that is degenerate at zero and a variant of the Poisson model. Conversely, the ZINB model is a mix of zero and a variant of negative binomial model.

Opposite situations from the zero-inflated models are also encountered; this article examines processes that have no zeroes: the zero-truncated models. If the Poisson or the negative binomial model is used with these processes, the procedure tries to fit the model by including probabilities for zero values. More accurate models that do not include zero values should be able to be produced. If the value of zero cannot be observed in any random experiment, then these models may be used. Two cases are considered: (1) the zero-truncated Poisson model, and (2) the zero-truncated negative binomial model.

Given a range of possible models, it is difficult to fit an appropriate discrete model. The main purpose of this article is to provide guidelines to fit a discrete process appropriately. First, a simulation study was conducted to determine the performance of the considered models when excess zeroes are present in a dataset. Second, the following real-life data (For details, see Table 4.1) were analyzed, the numbers of:

1. Road accidents per month in the Dhaka district,
2. People visiting the Dhaka medical hospital (BMSSU) per day,
3. Earthquakes in Bangladesh per year, and
4. Strikes (hartals) per month in Dhaka.

Statistical Distribution: The Binomial Distribution

If $X \sim B(n, p)$, then the probability mass function (pmf) of X is defined by

$$P(X=k;n,p)=n_{C_K}\,p^k(1-p)^{n-k},\quad k=0,1,2,\ldots,n. \qquad (2.1)$$

where n is the total number of trials and p is the probability of success of each trial. The moment generating function (mgf) of (2.1) is

$$M_X(t)=(p+qe^t)^n,$$

thus, E(X), V(X) and skewness (S_k) of (2.1) are np, npq and $[(1-2p)^2/npq]$ respectively.

Statistical Distribution: The Poisson Distribution

In (2.1) if $n\to\infty$ and $p\to 0$, then X follows the Poisson distribution with parameter $\lambda(>0)$ (denoted $X \sim P(\lambda)$). The pmf is defined as

$$P(X=k;\lambda)=\frac{e^{-\lambda}\lambda^k}{k!},\ k=0,1,2,\ldots \qquad (2.2)$$

where λ denotes expected number of occurrences. The mgf of (2.2) is

$$M_X(t)=e^{\lambda(e^t-1)},$$

thus, E(X) and V(X) of (2.2) are the same, which is λ and S_k is equal to $1/\lambda$.

Statistical Distribution: The Negative Binomial Distribution

If X~NB(k,p), then the pmf of X is given by

$$P(X;k,p)=\frac{\Gamma(k+X)}{X!\Gamma k}p^{X}q^{-(k+X)},\quad p>0,\ k>0,\ X=0,\ 1,\ 2,... \tag{2.3}$$

where p is the chance of success in a single trial and k are the number of failures of repeated identical trials. If $k\to\infty$, then X~P(λ), where $\lambda = kp$. The mgf of (2.3) is

$$M_X(t)=(q-pe^{t})^{-k},$$

thus, E(X), V(X) and S_k of (2.3) are kp, kpq and $[(1+2p)^2/kpq]$ respectively.

Statistical Distribution: The Zero-Inflated Poisson (ZIP) Distribution

If X~ZIP(θ, λ) with parameters θ and λ, then the pmf is defined by

$$P(X;\theta,\lambda)=\begin{cases}1-\theta & if\ X=0\\ \theta\dfrac{P(X;\lambda)}{1-e^{-\lambda}} & if\ X>0\end{cases} \tag{2.4}$$

where P(X; λ) is defined in (2.2) and θ is the proportion of non-zero values of X. The mgf of (2.4) is

$$M_X(t)=(1-\theta)+\frac{\theta e^{-\lambda}}{1-e^{-\lambda}}(e^{\lambda e^{t}}-1),$$

thus, E(X), V(X) and S_k of (2.4) are

$$\frac{\theta\lambda}{1-e^{-\lambda}},\ \frac{\theta\lambda}{1-e^{-\lambda}}\left[\lambda+1-\frac{\theta\lambda}{1-e^{-\lambda}}\right]$$

and

$$\frac{\lambda^2+3\lambda+1-\left(\frac{\theta\lambda}{1-e^{-\lambda}}\right)\left[3\lambda+3-\left(\frac{\theta\lambda}{1-e^{-\lambda}}\right)\right]^2}{\left(\frac{\theta\lambda}{1-e^{-\lambda}}\right)\left[\lambda+1-\left(\frac{\theta\lambda}{1-e^{-\lambda}}\right)\right]^3}$$

respectively.

Statistical Distribution: Zero-Inflated Negative Binomial (ZINB) Distribution

If X~ZINB(θ, k, p), then the pmf of X is defined by

$$P(X;\theta,k,p)=\begin{cases}1-\theta & if\ X=0\\ \theta\dfrac{P(X;k,p)}{1-q^{-k}} & if\ X>0\end{cases} \tag{2.5}$$

where P(X; k, p) is defined in (2.3) and θ is the proportion of non-zero values of X. The mgf of (2.5) is

$$M_X(t)=(1-\theta)+\frac{\theta}{1-q^{-k}}[(q-pe^{t})^{-k}-1],$$

thus, E(X), V(X) and S_k of (2.5) are

$$\frac{\theta kp}{1-q^{-k}},\ \frac{\theta kp}{1-q^{-k}}\left[p(k+1)+1-\left(\frac{\theta kp}{1-q^{-k}}\right)\right]$$

and

$$\frac{(kp)^2+3kp+1-\left(\frac{\theta kp}{1-q^{-k}}\right)\left[3kp+3-\left(\frac{\theta kp}{1-q^{-k}}\right)\right]^2}{\left(\frac{\theta kp}{1-q^{-k}}\right)\left[kp+1-\left(\frac{\theta kp}{1-q^{-k}}\right)\right]^3}$$

respectively. As $k\to\infty$, ZINB (θ, k, p) ~ZIP(θ, λ), where $\lambda = kp$.

Statistical Distribution: Zero-Truncated Poisson (ZTP) Distribution

If X~ZTP(λ), then the pmf of X is given by

$$P(X=x\mid X>0)=\frac{P(X=x)}{P(X>0)}=\frac{P(X;\lambda)}{1-P(X=0)}=\frac{P(X;\lambda)}{(1-e^{-\lambda})}$$

for x = 1, 2, 3, …, where P(X; λ) is defined in (2.2). The mgf of this distribution is

$$M_X(t) = \frac{e^{-\lambda}}{1-e^{-\lambda}}(e^{\lambda e^t} - 1),$$

thus, E(X) and V(X) are $\lambda(1-e^{-\lambda})^{-1}$ and S_k is $(1-e^{-\lambda})(\lambda+3+\lambda^{-1}) - 3(\lambda+1) - (1-e^{-\lambda})^{-1}$.

Statistical Distribution: Zero-Truncated Negative Binomial (ZTNB) Distribution

If X~ZTNB(k, p), then the pmf of X is given by

$$P(X = x \mid X > 0) = \frac{P(X; k, p)}{1 - P(X = 0)} = \frac{P(X; k, p)}{1 - q^{-k}}$$

for x = 1, 2, 3, …, where P(X; k, p) is defined in (2.3). The mgf of this distribution is

$$M_X(t) = \frac{1}{1-q^{-k}}[(q - pe^t)^{-k} - 1],$$

thus, E(X), V(X) and S_k are

$$\frac{kp}{1-q^{-k}}, \frac{kp}{1-q^{-k}}\left[p(k+1) + 1 - \left(\frac{kp}{1-q^{-k}}\right)\right]$$

and

$$\frac{(kp)^2 + 3kp + 1 - \left(\frac{kp}{1-q^{-k}}\right)\left[3kp + 3 - \left(\frac{kp}{1-q^{-k}}\right)\right]^2}{\left(\frac{kp}{1-q^{-k}}\right)\left[kp + 1 - \left(\frac{kp}{1-q^{-k}}\right)\right]^3}$$

respectively.

Parameter Estimation

To estimate the parameters of the considered models, the most common methods are the method of moment estimation (MME) (Pearson, 1894) and the maximum likelihood estimation (MLE) method (Fisher, 1922). The latter method has been used extensively since in the early 1900s, due to its properties of being consistent, asymptotically normal and having minimum variances for large samples.

The Moment Estimation Method (MME)

Consider the k^{th} moments of a random variable X. By notation,

$$M_k = \sum_{i=1}^{n} \frac{X_i^k}{n} = E(X^k),\ k = 1, 2, 3, \ldots,$$

thus,

$$M_1 = E(X),\ M_2 = \sum_{i=1}^{n} \frac{X_i^2}{n}, M_3 = \sum_{i=1}^{n} \frac{X_i^3}{n}.$$

The Maximum Likelihood Estimation Method (MLE)

Find the log-likelihood function for a given distribution and take a partial derivative of this function with respect to each parameter and set it equal to 0; solve it to find the parameters estimate.

Binomial Distribution: Moment Estimator of p

Based on (2.1), it is known that E(X) = np, therefore, $M_1 = np$. Simplification results in:

$$B_{\hat{p}(mom)} = E(X)/n = M_1/n.$$

Binomial Distribution: Maximum Likelihood Estimator of p

The log-likelihood expression of (2.1) is

$$LogL(X;n,p) = Constant + \sum_{i=1}^{n} X_i \log p + n - \sum_{i=1}^{n} X_i \log(1-p);$$

differentiating the above expression with respect to p, the following equation is obtained

$$\frac{\partial \log L(X; n, p)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_{i=1}^{p} X_i}{1-p}.$$

Simplifying results in:

$$B_{\hat{p}(ml)} = \sum_{i=1}^{n} X_i / n$$

.

Poisson Distribution: Moment Estimator of λ

Based on (2.2), it is known that E(X) = λ, thus,

$$P_{\hat{\lambda}(mom)} = M_1.$$

Poisson Distribution: Maximum Likelihood Estimator of λ

The log-likelihood expression of (2.2) is

$$LogL(X;\lambda) = -n\lambda + \sum_{i=1}^{n} X_i \log \lambda - \sum_{i=1}^{n} (\log X_i)!$$

Differentiating the above expression with respect to λ, results in

$$\frac{\partial \log L(X;\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^{p} X_i}{\lambda}$$

and, after simplification,

$$P_{\hat{\lambda}(ml)} = M_1$$

thus

$$P_{\hat{\lambda}(mom)} = P_{\hat{\lambda}(ml)} = M_1.$$

Negative Binomial Distribution: Moment Estimators of p and k

Based on (2.3), it is known that E(X) = kp and V(X) = kpq, thus

$$M_1 = kp \qquad (2.6)$$

and

$$M_2 - M_1^2 = kpq \qquad (2.7)$$

Solving (2.7) for q results in $\hat{q} = \frac{M_2 - M_1^2}{M_1}$, and because it is known (based on 2.3) that $q - p = 1$,

$$NB_{\hat{p}(mom)} = (s^2 / M_1) - 1.$$

Solving (2.6), results in

$$NB_{\hat{k}(mom)} = \frac{M_1^2}{s^2 - M_1}$$

where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_1)^2 .$$

Negative Binomial Distribution: Maximum Likelihood Estimators of p and k

The log-likelihood expression of (2.3) is

$$LogL(X; k, p) = \sum_{i=1}^{n} \log\left(\frac{\Gamma(X_i + k)}{X_i!\,\Gamma k}\right) + \sum_{i=1}^{n} X_i \log p - \sum_{i=1}^{n} (k + X_i) \log q$$

Differentiating the above expression with respect to p and k, the following equations result:

$$\frac{\partial LogL(X;k,p)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{\sum_{i=1}^{n} (k + X_i)}{1+p} \qquad (2.8)$$

and

$$\frac{\partial LogL(X;k,p)}{\partial k} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{X_i - 1} \frac{j}{1 + jk^{-1}} + k^2 \ln(1+p) - \frac{kp(E(X)+k)}{1+p} \qquad (2.9)$$

Solving (2.8), results in $NB_{p(ml)} = M_1 / \hat{k}$. It was observed that $NB_{\hat{k}(ml)}$ does not exist in closed form, thus, $NB_{\hat{k}(ml)}$ was obtained by optimizing numerically (2.9) using the Newton-Raphson optimization technique where $p = \hat{p}$.

ZIP Distribution: Moment Estimators of θ and λ

It is known for (2.4) that

$$E(X) = \frac{\theta\lambda}{1 - e^{-\lambda}}$$

and

$$V(X) = \frac{\theta\lambda}{1 - e^{-\lambda}} \left[\lambda + 1 - \frac{\theta\lambda}{1 - e^{-\lambda}} \right],$$

thus,

$$M_1 = \frac{\theta\lambda}{1 - e^{-\lambda}},$$

and

$$M_2 - M_1^2 = \frac{\theta\lambda}{1 - e^{-\lambda}} \left[\lambda + 1 - \frac{\theta\lambda}{1 - e^{-\lambda}} \right].$$

Simplifying the above equations results in

$$ZIP_{\hat{\lambda}(mom)} = \frac{M_2}{M_1} - 1$$

and

$$ZIP_{\hat{\theta}(ml)} = \frac{M_1(1-e^{-\hat{\lambda}})}{\hat{\lambda}}.$$

ZIP Distribution: Maximum Likelihood Estimators of θ and λ

The log-likelihood expression of (2.4) is

$$LogL(X;\ \theta,\ \lambda) = \sum_{i=1}^{n} I(X_i = 0)\log(1-\theta) + \sum_{i=1}^{n} I(X_i > 0)\{\log\theta + X_i \log\lambda - \log(e^{\lambda}-1) - \log X_i!\}$$

Differentiating the above expression with respect to θ and λ, results in

$$\frac{\partial\, log\, L(X;\theta,\lambda)}{\partial\theta} = -\frac{\sum_{i=1}^{n} I(X_i = 0)}{1-\theta} + \frac{\sum_{i=1}^{n} I(X_i > 0)}{\theta}$$

$$\frac{\partial \log \mathrm{L}(\mathrm{X};\theta,\lambda)}{\partial\lambda} = \sum_{i=1}^{n} \mathrm{X}_i / \lambda - \sum_{i=1}^{n} \mathrm{I}(\mathrm{X}_i > 0) - \sum_{i=1}^{n} \mathrm{I}(\mathrm{X}_i > 0)\frac{e^{-\lambda}}{e^{\lambda}-1}$$

After the above equations are simplified for θ and λ, the following are obtained:

$$ZIP_{\hat{\theta}(mom)} = \sum_{i=1}^{n} I(X_i > 0)/n$$

and

$$ZIP_{\hat{\lambda}(ml)} = E(X)(1-e^{-\lambda}).$$

where E(X) is the expected value of the non-zero occurrences of X (λ does not have a closed form solution, hence the Newton-Raphson algorithm was used to find $\hat{\lambda}$ iteratively.)

ZINB Distribution: Moment Estimators of θ, k, p

Moment estimators of θ, k, p do not exist.

ZINB Distribution: Maximum Likelihood Estimators of θ, k, p

The log-likelihood expression of (2.5) is

$$LogL(X;\theta,k,p) = \sum_{i=1}^{n} I(X_i = 0)\log(1-\theta) + \sum_{i=1}^{n} I(X_i > 0)\log\theta + \sum_{i=1}^{n} \log\left(\frac{\Gamma(X_i+K)}{X_i!\Gamma K}\right) - \sum_{i=1}^{n} (k+X_i)\log q + \sum_{i=1}^{n} X_i \log p - \sum_{i=1}^{n} \log(1-q^{-k})$$

Differentiating the above with respect to each of parameters, results in the following estimators for θ, k, and p:

$$ZINB_{\hat{\theta}(ml)} = \sum_{i=1}^{n} I(X_i > 0)/n.$$

Other estimates $\hat{p}$ *and* $\hat{k}$ were found iteratively: k, p is given by

$$ZINB_{\hat{p}(ml)} = E(X)\{1-(1+kp)^{-k^{-1}}\} \quad (2.10)$$

thus, the solution of $\hat{p}$ has the same properties as described above. Because the score equation for $\hat{k}$ does not have a simple form, k was estimated numerically given the current estimate of $\hat{p}$ from (2.10) (for details, see Warton, 2005).

ZTP Distribution

The estimated parameters $ZTP_{\hat{\lambda}(mom)}$ and $ZTP_{\hat{\lambda}(ml)}$ are similar to $ZIP_{\hat{\lambda}(mom)}$ and $ZIP_{\hat{\lambda}(ml)}$, where the log-likelihood expression for this distribution is given by

$$LogL(X;\ \lambda) = \sum_{i=1}^{n} I(X_i > 0)\{X_i\ log\ \lambda - log(e^{\lambda}-1) - log\ X_i!\}$$

.

ZTNB Distribution

The estimated parameters $ZTNB_{\hat{p}(ml)}$ and $ZTNB_{\hat{k}(ml)}$ are similar to $ZINB_{\hat{p}(ml)}$ and $ZINB_{\hat{k}(ml)}$, where the log-likelihood expression for this distribution is given by

$$LogL(X;\ k,\ p) =$$

$$\sum_{i=1}^{n}\log(\frac{\Gamma(X_i+k)}{X_i!\Gamma k})+\sum_{i=1}^{n}X_i\log p$$
$$-\sum_{i=1}^{n}(k+X_i)\log q-\sum_{i=1}^{n}\log(1-q^{-k})$$

Methods for Comparison of the Distributions:
Goodness of Fit (GOF) Test

The GOF test determines whether a hypothesized distribution can be used as a model for a particular population of interest. Common tests include the χ^2, the Kolmogorov-Smirnov and the Anderson-Darling tests. The χ^2 test can be applied for discrete models; other tests tend to be restricted to continuous models. The test procedure for the χ^2 GOF is simple: divide a set of data into a number of bins and the number of points that fall into each bin is compared to the expected number of points for those bins (if the data are obtained from the hypothesized distribution). More formally, suppose:

H_0: Data follows the specified population distribution.
H_1: Data does not follow the specified population distribution.

If the data is divided into bins, then the test statistic is:

$$\chi^2_{cal}=\sum_{i=1}^{s}\frac{(O_i-E_i)^2}{E_i} \qquad (2.11)$$

where O_i and E_i are the observed and expected frequencies for bin i. The null, H_o, is rejected if $\chi^2_{cal}>\chi^2_{df,\alpha}$, where degrees of freedom (df) is calculated as (s−1− # of parameters estimated) and α is the significance level.

Methodology

Simulation Study

Because the outcome of interest in many fields is discrete in nature and generally follows the binomial, Poisson or the negative binomial distribution. It is evident from the literature that these types of variables often contain a high proportion of zeroes. These zeroes may be due to either the presence of a population with only zero counts and/or over-dispersion. Hence, it may be stated that - to capture the effect of excess zeroes - it is necessary to investigate which model would best fit a discrete process. Thus, a series of simulation experiments was conducted to determine the effect of excess zeroes on selected models. These simulation studies reflect how commonly used discrete models behave if excess zeroes are present in a set of data.

Simulation Experiment Design

A sample, $X = \{X_1, X_2, \ldots, X_n\}$, was obtained where data were generated from a Poisson model with:

λ: 1.0, 1.5, 2.0 and 2.5;
n: 10, 20, 30, 50, 100, 150, 200; and
10%, 20%, 80% zeroes.

Different data sets for different sample sizes and λs were generated to determine which model performs best if zeroes (10% to 80%) are present in a dataset. To select the possible best model, the Chi-square GOF statistic defined in (2.11) for all models were calculated. If the test was not statistically significant, then the data follows the specified (e.g., binomial, Poisson, or other) population distribution. Tables 3.1-3.8 show the GOF statistic values for all proposed distributions. Both small and large sample behaviors were investigated for all models, and all calculations were carried out using the programming code MATLAB (Version 7.0).

Results

Tables 3.1 to 3.8 show that the performance of the models depends on the sample size (n), λ and the percentage of zeroes included in the sample. It was observed that, as the percentage of zeroes increases in the sample, the proportion of over dispersion decreases and performance of the binomial and Poisson distributions decrease. For small sample sizes, most of the models fit well; for large sample sizes, however, both the binomial and Poisson performed poorly compared to others. For samples containing a moderate to high percentage of zeroes, the negative binomial performed best followed by ZIP and ZINB. Based on simulations, therefore, in the presence of excess zeroes the negative binomial model and the ZIP (moment estimator of parameters) model to approximate a real discrete process are recommended.

Table 3.1: Simulation Results for 10% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(1.87, 1.55)	4.86	1.53	0.71	-	-	1.45	0.93	1.08
	1.5	(1.88, 1.61)	4.59	1.45	0.49	-	-	0.80	0.61	0.48
	2.0	(2.11, 1.11)	6.99	3.24	3.94	-	-	1.26	4.54	6.34
	2.5	(2.66, 2.00)	2.41	14.34	2.11	-	-	-	1.32	1.96
20	1.0	(1.29, 0.59)	3.14	1.03	3.56	-	-	0.87	0.22	57.69*
	1.5	(1.55, 1.67)	50.39*	22.04*	6.80	6.57	6.74	6.95	6.00	4.35
	2.0	(1.89, 1.43)	8.03	2.07	0.64	-	-	1.70	0.75	0.70
	2.5	(2.42, 2.47)	33.68*	15.00	4.53	4.53	4.43	4.68	5.15	6.19
30	1.0	(1.28, 0.71)	4.74	1.22	3.04	-	-	0.85	0.95	111.05*
	1.5	(1.78, 1.45)	10.13	3.22	1.16	-	-	1.67	1.55	1.03
	2.0	(2.11, 2.41)	8.48	31.46*	11.36	9.80	10.08	11.81	7.60	8.91
	2.5	(2.25, 2.66)	129.94*	51.96*	6.42	5.15	5.33	6.01	8.45	4.31
50	1.0	(1.50, 1.00)	21.9	6.37	5.58	-	-	5.97	5.03	4.76
	1.5	(1.82, 1.25)	10.75	2.09	1.73	-	-	6.93	2.28	4.78
	2.0	(2.33, 2.04)	57.09*	24.38*	6.99	-	-	10.67	8.55	9.52
	2.5	(2.68, 2.21)	34.43*	26.04*	10.41	-	-	50.46*	11.36	16.67
100	1.0	(1.24, 0.54)	13.33	4.32	15.87	-	-	5.06	2.04	217.25*
	1.5	(1.67, 1.29)	32.99*	11.59	6.36	-	-	2.50	2.63	7.72
	2.0	(1.93, 1.74)	73.11*	27.26*	4.50	-	-	3.37	4.37	1.79
	2.5	(2.50, 3.27)	379.60*	159.59*	9.64	4.74	4.78	11.31	12.02	7.29
150	1.0	(1.21, 0.64)	27.15*	18.09*	28.80*	-	-	1.10	2.05	630.30*
	1.5	(1.68, 1.35)	33.19*	11.91	6.94	-	-	1.51	1.83	9.04
	2.0	(2.50, 2.49)	108.42*	51.92*	5.80	-	-	6.54	7.28	17.72*
	2.5	(2.05, 1.80)	54.72*	19.62*	2.98	-	-	3.44	3.86	9.24
200	1.0	(1.31, 0.87)	33.01*	20.37*	25.95*	-	-	3.81	3.53	48.64*
	1.5	(1.76, 1.58)	281.53*	111.60*	11.22	-	-	8.43	10.59	18.00*
	2.0	(2.16, 2.16)	98.78*	39.66*	0.52	-	-	0.51	2.67	9.86
	2.5	(2.52, 2.49)	108.42*	51.92*	5.80	-	-	6.54	7.28	17.72*

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.2: Simulation Results for 20% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.75, 0.91)	2.44	0.87	0.54	0.56	0.54	0.67	1.49	19.74*
	1.5	(1.88, 1.61)	1.90	1.36	0.21	0.33	0.23	0.11	0.65	0.44
	2.0	(2.11, 1.11)	5.13	1.97	1.07	-	-	2.51	2.08	1.78
	2.5	(2.40, 2.26)	4.46	52.17*	2.16	-	-	5.88	1.24	2.33
20	1.0	(1.14, 0.74)	9.12	2.43	3.54	-	-	5.78	5.42	83.15*
	1.5	(1.76, 1.69)	9.47	6.16	3.42	-	-	4.59	5.00	3.93
	2.0	(2.05, 2.26)	43.37*	21.00*	6.54	6.17	6.30	6.01	7.87	7.19
	2.5	(2.25, 2.72)	22.62*	17.07*	7.25	6.46	6.41	5.08	7.77	7.53
30	1.0	(1.44, 1.11)	16.47	4.57	1.98	4.14	-	-	4.42	1.94
	1.5	(1.51, 1.87)	15.05	6.69	2.80	3.36	3.13	3.70	3.85	3.97
	2.0	(1.71, 3.02)	183.02*	79.97*	10.86	2.53	2.23	3.38	17.55*	1.50
	2.5	(1.89, 1.87)	22.97*	14.24	6.21	-	-	7.73	9.48	8.64
50	1.0	(1.07, 0.73)	8.91	1.90	3.01	-	-	0.72	0.87	226.10*
	1.5	(1.27, 1.12)	9.23	2.64	0.84	-	-	0.82	2.45	2.57
	2.0	(1.42, 1.58)	36.468	14.27	0.83	0.40	0.45	0.74	4.39	0.90
	2.5	(2.31, 4.26)	1.16e+003*	473.03*	25.04*	9.44	9.44	17.75*	29.23*	21.43*
100	1.0	(1.27, 1.11)	12.51	4.70	2.93	-	-	5.07	8.63	6.35
	1.5	(1.54, 1.98)	111.79*	44.34*	1.81	1.08	1.13	1.68	8.20	3.94
	2.0	(1.90, 2.02)	49.98*	29.58*	10.09	9.46	9.43	7.80	13.15	12.44
	2.5	(2.21, 2.95)	129.95*	73.67*	17.07*	10.76	10.42	3.83	13.62	11.96
150	1.0	(1.15, 0.92)	26.42*	6.98	4.04	-	-	0.79	3.19	15.99
	1.5	(1.38, 1.45)	521.78*	206.30*	13.68	11.50	11.97	13.60	37.08*	11.87
	2.0	(2.09, 2.85)	643.42*	284.48*	27.75*	9.65	8.65	4.70	31.70*	9.87
	2.5	(1.81, 2.35)	163.00*	77.96*	10.29	2.76	2.33	2.08	17.60*	2.32
200	1.0	(1.07, 0.80)	16.68	2.76	5.02	-	-	0.60	2.64	931.68*
	1.5	(1.43, 1.39)	31.52*	12.61	2.07	-	-	2.76	12.21	7.39
	2.0	(1.78, 2.20)	269.23*	119.66*	12.80	7.57	7.52	6.05	23.29*	12.43
	2.5	(2.09, 2.85)	643.42*	284.48*	27.75*	9.65	8.65	4.70	31.71*	9.87

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.3: Simulation Results for 30% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.87, 0.69)	2.84	0.45	0.44	-	-	0.94	1.16	27.88*
	1.5	(1.12, 1.83)	3.96	2.01	0.55	0.55	0.44	0.91	1.43	0.67
	2.0	(2.10, 2.98)	5.61	14.43	3.24	1.23	1.09	1.51	0.49	0.36
	2.5	(2.30, 4.90)	18.52*	19.26*	6.62	3.92	4.02	1.51	3.76	2.96
20	1.0	(0.94, 0.71)	3.21	0.51	0.86	-	-	0.15	0.39	92.09*
	1.5	(1.38, 1.78)	35.44*	15.61	4.36	3.13	3.13	2.60	9.45	2.50
	2.0	(1.70, 2.09)	20.71*	17.90*	11.35	10.73	10.68	8.74	13.98	11.83
	2.5	(1.65, 2.39)	17.36*	11.52	4.39	3.82	3.82	2.42	2.03	4.38
30	1.0	(1.0, 0.88)	7.18	1.64	0.16	-	-	0.50	2.29	190.49*
	1.5	(1.0, 0.91)	7.71	2.04	0.18	-	-	0.59	2.85	197.58*
	2.0	(2.03, 3.89)	32.69*	52.35*	21.40*	6.51	3.97	6.44	4.58	3.54
	2.5	(1.29, 2.21)	80.02*	33.24*	4.13	1.99	1.78	5.92	12.27	4.58
50	1.0	(0.84, 0.97)	6.60	3.12	1.70	1.69	1.67	1.23	5.90	3.52
	1.5	(1.48, 2.30)	283.18*	120.03*	13.83	5.50	5.26	2.89	29.87*	4.96
	2.0	(1.67, 2.09)	64.25*	35.99*	11.66	9.08	8.83	6.04	14.96	9.07
	2.5	(2.02, 3.32)	551.89*	241.91*	37.23*	8.70	7.04	1.51	104.44*	7.85
100	1.0	(0.92, 0.84)	12.54	3.18	0.59	-	-	0.59	5.54	663.75*
	1.5	(1.23, 1.72)	156.60*	67.78*	9.01	1.21	0.95	1.41	34.08*	0.48
	2.0	(1.47, 2.04)	146.74*	69.81*	13.45	4.26	3.48	1.25	28.87*	1.80
	2.5	(1.92, 3.14)	2.62e+003*	1.06e+003*	51.79*	21.29*	20.20*	7.95	66.55*	20.26*
150	1.0	(0.93, 1.00)	37.20*	13.93	0.89	0.28	0.29	0.25	17.34*	11.65
	1.5	(1.27, 1.50)	46.01*	26.46*	10.28	8.17	8.05	4.44	27.61*	9.47
	2.0	(1.82, 2.79)	532.08*	273.65*	61.18*	26.09*	22.16*	4.35	64.56*	20.42*
	2.5	(1.44, 2.07)	281.77*	127.63*	22.52*	5.78	4.76	6.99	54.01*	2.57
200	1.0	(0.95, 0.99)	72.01*	27.23*	1.18	0.83	0.86	1.22	22.25*	16.74
	1.5	(1.15, 1.47)	88.14*	44.03*	10.55	4.57	4.41	2.79	44.02*	5.86
	2.0	(1.61, 2.49)	1.75e+003*	723.43*	48.01*	12.80	11.61	3.90	92.98*	10.08
	2.5	(1.82, 2.79)	532.08*	273.65*	61.18*	26.098*	22.16*	4.35	64.56*	20.42*

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.4: Simulation Results for 40% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.66, 1.46)	2.26	1.87	0.96	0.53	0.58	0.17	1.86	0.66
	1.5	(1.20, 1.73)	4.03	3.02	1.38	1.31	1.28	0.67	2.61	1.46
	2.0	(1.40, 2.04)	17.73*	9.68	3.77	2.71	2.68	1.85	6.30	2.68
	2.5	(1.90, 3.21)	10.75	27.79*	6.48	3.99	3.90	2.16	3.84	4.04
20	1.0	(0.83, 0.97)	21.23*	8.40	3.17	2.69	2.79	2.85	9.66	2.70
	1.5	(0.84, 0.69)	8.90	2.35	1.63	-	-	2.89	4.66	79.64*
	2.0	(1.07, 2.07)	42.96*	22.27*	8.79	3.94	3.62	1.38	24.15*	2.86
	2.5	(1.70, 3.58)	40.06*	26.74*	9.49	3.32	3.24	0.34	8.61	1.42
30	1.0	(0.76, 1.02)	75.62*	31.05*	6.16	3.57	3.64	5.07	28.62*	4.84
	1.5	(0.92, 1.27)	35.83*	15.13	2.78	0.62	0.62	0.95	14.79	0.71
	2.0	(1.41, 2.89)	267.73*	126.08*	47.17*	27.79*	28.18*	13.98	103.35*	35.84*
	2.5	(1.36, 2.30)	81.23*	39.25*	9.82	2.59	2.16	1.75	22.33*	1.91
50	1.0	(1.04, 1.57)	150.54*	63.48*	8.05	1.16	0.99	0.98	39.99*	0.59
	1.5	(1.17, 1.65)	70.74*	33.93*	9.43	4.40	3.46	1.46	26.17*	1.39
	2.0	(1.51, 2.50)	200.58*	95.75*	22.46*	8.65	7.25	2.60	38.05*	5.88
	2.5	(1.34, 2.36)	115.07*	62.35*	23.13*	8.98	7.79	4.56	50.28*	8.02
100	1.0	(0.87, 1.29)	232.718	99.50*	14.12	3.43	3.24	2.03	88.05*	3.19
	1.5	(1.06, 1.70)	303.13*	134.46*	35.92*	14.85	12.96	7.92	119.83*	10.78
	2.0	(1.26, 1.87)	144.32*	74.04*	21.62*	9.68	8.51	2.94	50.29*	6.50
	2.5	(1.51, 3.11)	1.78e+003*	742.21*	77.04*	11.73	8.38	4.45	181.85*	4.43
150	1.0	(0.79, 0.91)	33.33*	14.98	3.49	1.57	1.44	0.68	28.07*	1.4e+003*
	1.5	(1.03, 1.62)	795.95*	330.68*	27.73*	4.15	3.65	1.85	160.24*	3.04
	2.0	(1.60, 3.65)	2.18e+004*	8.6e+003*	221.00*	20.44*	15.72	9.98	623.30*	5.46
	2.5	(1.31, 2.38)	1.2e+003*	571.42*	81.64*	18.53*	14.69	12.03	232.37*	16.67
200	1.0	(0.75, 1.00)	206.88*	87.74*	11.31	2.48	2.36	2.94	91.17*	7.66
	1.5	(1.06, 1.55)	210.87*	103.49*	25.11*	6.13	4.77	0.47	95.82*	2.48
	2.0	(1.24, 2.04)	1.68e+003*	699.68*	62.48*	13.81	12.28	5.15	225.91*	10.27
	2.5	(1.60, 3.65)	2.18e+004*	8.69e+003*	221.00*	20.44*	15.72	9.98	623.30*	5.46

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.5: Simulation Results for 50% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.60, 0.48)	1.40	0.02	0.19	-	-	0.13	0.16	17.86*
	1.5	(1.10, 2.10)	43.97*	19.56*	5.32	1.60	1.47	0.94	17.22*	1.33
	2.0	(1.00, 1.75)	27.08*	13.12	6.53	4.48	4.46	2.63	17.72*	4.57
	2.5	(0.62, 0.83)	8.88	4.02	2.78	2.54	2.54	2.08	6.72	62.37*
20	1.0	(0.66, 0.82)	10.80	4.24	0.93	0.29	0.32	0.28	5.81	241.12*
	1.5	(0.56, 0.79)	20.75*	8.55	2.43	0.77	0.80	0.81	11.44	0.66
	2.0	(0.85, 1.29)	59.76*	24.70*	4.43	0.95	0.99	1.61	24.03*	0.93
	2.5	(1.21, 2.50)	169.46*	74.33*	13.89	3.18	2.60	0.72	44.27*	3.36
30	1.0	(0.42, 0.55)	16.33	6.96	3.94	2.75	2.59	1.97	10.81	121.28*
	1.5	(0.85, 1.51)	74.89*	73.14*	14.97	3.17	2.85	2.90	78.40*	2.11
	2.0	(1.03, 2.24)	232.29*	113.74*	26.09*	9.27	8.85	3.24	106.43*	9.77
	2.5	(1.41, 3.10)	225.50*	107.76*	28.80*	9.25	7.93	2.21	60.34*	11.61
50	1.0	(0.64, 0.91)	89.04*	37.10*	6.18	0.95	0.89	1.01	42.96*	1.12
	1.5	(0.89, 1.61)	42.06*	22.55*	8.57	2.36	2.33	2.33	30.61*	1.65
	2.0	(0.92, 2.06)	1.09e+003*	448.56*	51.15*	5.58	4.49	4.26	371.95*	2.83
	2.5	(1.30, 2.75)	551.89*	241.91*	37.23*	8.70	7.04	1.51	104.44*	7.85
100	1.0	(0.57, 0.82)	107.32*	46.20*	14.72	4.27	3.34	2.21	66.71*	0.87
	1.5	(0.78, 1.17)	284.75*	118.19*	20.75*	3.50	3.12	5.04	124.37*	2.16
	2.0	(1.10, 2.27)	822.90*	358.46*	49.51*	10.05	8.64	0.52	204.08*	7.83
	2.5	(1.25, 2.68)	577.43*	281.76*	68.86*	15.87*	11.79	1.33	195.13*	15.77
150	1.0	(0.61, 0.83)	94.65*	44.15*	17.37*	8.16	6.78	3.50	68.03*	1.8e+003*
	1.5	(0.91, 1.55)	877.63*	371.87*	52.36*	11.54	9.98	3.97	305.59*	7.01
	2.0	(1.28, 2.68)	8.3e+003*	3.3e+003*	179.29*	38.23*	35.12*	8.73	758.37*	38.32*
	2.5	(1.16, 2.43)	1.6e+003*	713.45*	94.96*	12.30	8.31	1.29	386.70*	6.23
200	1.0	(0.65, 0.90)	116.46*	55.23*	16.14	4.99	4.20	1.02	85.16*	7.88
	1.5	(0.82, 1.30)	210.40*	104.89*	29.59*	6.11	5.12	0.14	147.99*	2.77
	2.0	(1.23, 2.95)	4.4e+004*	1.7e+004*	312.14*	13.44	10.50	8.88	2.4e+003*	5.87
	2.5	(1.28, 2.68)	8.3e+003*	3.3e+003*	179.29*	38.23*	35.12*	8.73	758.37*	38.32*

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.6: Simulation Results for 60% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.44, 0.52)	3.33	1.14	0.45	0.29	0.36	0.33	1.73	31.31*
	1.5	(0.60, 0.93)	0.93	0.60	0.07	0.17	0.08	0.30	0.42	0.70
	2.0	(0.80, 1.73)	11.61	5.97	2.33	0.59	0.59	0.43	7.49	0.50
	2.5	(0.77, 1.69)	5.19	4.88	2.81	2.01	2.48	0.85	4.85	2.08
20	1.0	(0.35, 0.36)	2.66	0.61	0.09	0.06	0.09	0.13	0.91	36.26*
	1.5	(0.57, 0.81)	17.04*	7.31	2.20	0.56	0.50	0.29	10.51	0.13
	2.0	(0.95, 2.26)	522.66*	215.30*	25.74*	2.22	1.70	1.57	181.72*	0.61
	2.5	(1.36, 4.13)	102.70*	58.03*	20.07*	3.54	2.50	1.39	29.57*	4.79
30	1.0	(0.48, 0.56)	16.46	6.57	3.45	2.69	2.58	2.17	10.16	129.52*
	1.5	(0.79, 1.31)	70.10*	33.22*	14.84	8.04	7.31	3.93	50.71*	6.55
	2.0	(0.99, 1.98)	550.56*	116.98*	30.89*	9.87	8.45	2.89	98.09*	9.03
	2.5	(1.00, 2.59)	100.29*	464.03*	56.83*	7.57	6.21	2.37	407.44*	5.59
50	1.0	(0.37, 0.38)	8.58	2.31	0.53	0.44	0.52	0.52	3.41	104.71*
	1.5	(0.55, 1.30)	155.03*	66.52*	12.75	0.66	0.65	3.41	89.77*	2.93
	2.0	(0.91, 2.12)	370.66*	171.86*	44.30*	12.42	11.59	5.35	230.82*	12.37
	2.5	(1.12, 2.94)	4.8e+003*	1.95e+003*	118.24*	10.88	8.87	3.87	893.32*	5.71
100	1.0	(0.52, 0.87)	2.1e+003*	862.93*	59.97*	1.48	1.60	6.54	796.85*	4.73
	1.5	(0.57, 1.27)	1.7e+003*	729.47*	89.54*	4.64	3.85	7.12	1.01e+003*	3.43
	2.0	(0.82, 1.96)	4.2e+003*	1.75e+003*	121.95*	9.62	7.23	4.40	1.47e+003*	3.86
	2.5	(1.09, 2.79)	1.5e+003*	718.16*	136.83*	21.88*	18.27*	5.12	747.44*	14.69
150	1.0	(0.41, 0.50)	64.26*	25.91*	5.86	0.79	0.63	0.71	35.15*	2.02e+003*
	1.5	(0.68, 1.25)	426.14*	196.29*	58.20*	12.82	9.72	3.99	306.42*	7.49
	2.0	(1.07, 2.72)	9.1e+003*	3.87e+003*	298.17*	34.00*	27.98*	4.46	3.27e+003*	8.91
	2.5	(0.81, 1.69)	2.5e+003*	1.06e+003*	120.97*	8.71	5.49	3.67	1.05e+003*	5.61
200	1.0	(0.51, 0.87)	882.24*	369.87*	63.28*	4.00	2.63	3.03	482.26*	1.13
	1.5	(0.55. 0.93)	965.77*	409.93*	55.95*	6.49	5.24	2.93	498.47*	3.36
	2.0	(0.86, 1.74)	650.33*	329.36*	110.74*	30.81*	25.08*	5.69	573.57*	21.21*
	2.5	(1.07, 2.72)	9.1e+003*	3.87e+003*	298.17*	34.20*	27.98*	4.46	2.37e+003*	18.91*

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.7: Simulation Results for 70% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.60, 1.15)	29.65*	13.37	5.27	1.86	1.71	0.83	22.54*	1.39
	1.5	(0.50, 0.72)	11.8	5.37	3.54	2.84	2.83	2.07	8.99	82.23*
	2.0	(0.90, 2.32)	10.40	8.74	4.82	1.22	1.18	0.11	7.89	1.36
	2.5	(0.80, 1.73)	4.99	6.31	3.49	1.88	2.47	0.71	4.64	51.86*
20	1.0	(0.36, 0.57)	1.80	0.85	0.21	0.13	0.07	0.42	0.82	1.68
	1.5	(0.27, 0.33)	4.68	1.65	0.60	0.22	0.24	0.20	2.22	52.16*
	2.0	(0.57, 1.25)	23.82*	13.34	7.45	3.91	4.25	1.51	23.59*	3.71
	2.5	(1.10, 3.98)	293.96*	133.96*	31.52*	6.20	6.52	0.91	149.41*	1.67
30	1.0	(0.39, 0.61)	39.03*	17.02*	5.44	0.83	0.62	0.29	25.18*	0.13
	1.5	(0.72, 2.06)	109.55*	54.26*	20.14*	5.12	5.25	0.85	101.27*	1.79
	2.0	(0.82, 2.29)	790.34*	335.52*	50.67*	7.50	6.50	1.20	490.86*	14.03
	2.5	(0.50, 0.74)	23.87*	11.81	5.61	2.66	2.38	1.22	19.68*	442.47*
50	1.0	(0.41, 0.78)	477.66*	198.32*	30.46*	1.24	0.87	0.91	267.62*	0.33
	1.5	(0.47, 0.92)	202.94*	88.76*	31.60*	8.81	7.51	3.60	157.73*	5.92
	2.0	(0.56, 1.14)	159.68*	71.58*	18.19	1.70	1.46	1.53	112.87*	0.87
	2.5	(0.87, 3.55)	1.29e+003*	542.57*	70.87*	6.72	7.13	8.29	675.26*	104.59*
100	1.0	(0.29, 0.39)	61.89*	26.49*	14.07	7.28	6.18	4.32	39.15*	486.00*
	1.5	(0.43, 0.77)	242.08*	110.12*	30.088	5.98	4.92	1.25	181.32*	2.70
	2.0	(0.71, 2.01)	1.1e+003*	4.62e+003*	290.16*	13.95	11.20	2.57	6.0e+003*	8.88
	2.5	(0.81, 2.12)	4.8e+003*	2.01e+003*	174.84*	23.30*	21.50*	5.88	2.3e+003*	29.29*
150	1.0	(0.40, 0.71)	311.81*	146.78*	46.22*	5.37	5.81	1.57	270.69*	1.45
	1.5	(0.49, 0.93)	396.50*	189.83*	60.48*	19.02*	17.47*	6.49	335.32*	13.90
	2.0	(0.73, 1.86)	7.0e+003*	2.92e+003*	327.54*	29.06*	21.54*	5.85	3.8e+003*	24.25*
	2.5	(0.57, 1.18)	749.89*	336.53*	89.94*	20.35*	16.90	4.70	589.06*	12.55
200	1.0	(0.29, 0.38)	75.97*	33.90*	11.71	1.88	1.36	0.40	50.56*	2.3e+003*
	1.5	(0.56, 1.34)	1.9e+004*	7.99e+003*	304.89*	13.25	11.95	2.23	8.0e+003*	6.22
	2.0	(0.73, 1.94)	4.6e+003*	1.98e+003*	227.50*	30.41*	27.78*	5.59	2.7e+003*	15.97
	2.5	(0.73, 1.86)	7.0e+003*	2.92e+003*	327.54*	29.06*	21.54*	5.85	3.8e+003*	24.35*

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Table 3.8: Simulation Results for 80% Zero-Inflation

n	λ	The Chi-Square GOF Statistic								
		(Mean, Var)	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{mom}	ZIP_{ml}	$ZINB_{ml}$
10	1.0	(0.22, 0.44)	0	0.64	0.31	0.03	0.28	0.10	0.0032	0
	1.5	(0.50, 1.16)	3.17	3.24	2.05	0.51	0.67	0.07	2.91	31.48*
	2.0	(0.60, 1.82)	2.81	4.41	2.74	0.44	0.79	0.46	2.21	0.40
	2.5	(0.50, 1.16)	3.17	3.24	2.05	0.51	0.67	0.07	2.91	31.48*
20	1.0	(0.31, 0.67)	108.52*	46.91*	14.89	2.32	1.98	0.82	86.19*	1.54
	1.5	(0.40, 1.30)	1.80	2.65	1.54	0.56	0.39	8.46	1.17	64.35*
	2.0	(0.70, 2.32)	139.56*	64.29*	18.14*	2.61	2.33	0.16	134.54*	6.49
	2.5	(0.65, 2.55)	184.74*	81.78*	20.07*	2.28	3.38	0.94	173.58*	10.92
30	1.0	(0.11, 0.18)	24.48*	10.85	5.71	1.27	1.09	0.65	15.83	208.91*
	1.5	(0.34, 1.01)	53.98*	24.48*	8.24	0.46	0.47	3.31	41.92*	9.42
	2.0	(0.46, 1.29)	452.56*	189.90*	33.05*	2.09	1.79	0.32	337.42*	0.36
	2.5	(0.31, 0.57)	79.11*	35.05*	12.03	2.12	1.70	0.64	59.23*	0.94
50	1.0	(0.19, 0.29)	41.41*	18.28*	9.89	3.64	3.10	1.96	27.61*	328.00*
	1.5	(0.39, 1.05)	580.55*	247.89*	47.62*	4.17	3.73	0.27	481.91*	2.36
	2.0	(0.42, 0.82)	167.33*	76.05*	29.76*	9.56	8.86	3.91	144.92*	4.1e+003*
	2.5	(0.38, 0.85)	38.27*	20.88*	11.31	3.89	4.04	1.42	37.61*	2.45
100	1.0	(0.22, 0.36)	247.15	105.64	29.37*	1.74	1.19	0.93	147.52*	7.6e+003*
	1.5	(0.29, 0.56)	863.01*	362.55*	57.75*	1.83	1.08	0.27	513.18*	0.10
	2.0	(0.44, 1.30)	1.86e+004*	7.6e+003*	540.37*	12.36	10.00	2.66	1.2e+004*	11.11
	2.5	(0.54, 1.83)	1.59e+006*	6.2e+005*	7.1e+003*	9.56	7.22	2.52	6.2e+005*	1.50
150	1.0	(0.23, 0.41)	365.82*	160.49*	51.41*	5.92	4.22	1.42	256.57*	9.9e+003*
	1.5	(0.29, 0.55)	755.46*	331.02*	71.96*	4.16	2.68	0.42	543.71*	0.52
	2.0	(0.45, 1.12)	9.3e+003*	3.8e+003*	369.57*	20.28*	16.55	3.03	6.4e+003*	26.67*
	2.5	(0.48, 1.32)	1.7e+004*	7.2e+003*	530.55*	17.27*	13.11	2.32	1.1e+004*	17.63*
200	1.0	(0.22, 0.41)	1.46e+003*	611.40*	86.64*	3.60	3.36	1.53	776.13*	2.75
	1.5	(0.36, 0.84)	6.9e+003*	2.8e+003*	280.71*	9.29	6.20	0.34	4.5e+003*	7.44
	2.0	(0.42, 1.30)	4.1e+005*	1.6e+005*	4.4e+003*	12.41	8.16	4.40	2.1e+005*	1.73
	2.5	(0.45, 1.12)	9.3e+003*	3.8e+003*	369.57*	20.28*	16.55	3.03	6.4e+003*	26.57*

Notes: $\chi^2_{9,0.05}$ =16.91; *indicates significance (data do not follow the mentioned distribution) at α = 5%; - indicates over-dispersion; mom denotes moment estimator, and ml denotes MLE of model parameters.

Applications to Real Data Sets

The selected processes were fitted using theoretical principles and by understanding the simulation outcomes. Theoretical explanations of a discrete process are reviewed as follows: Generally, a process with two outcomes (see Lord, et al., 2005, for details) follows a Bernoulli distribution. To be more specific, consider a random variable, which is NOA. Each time a vehicle enters any type of entity (a trial) on a given transportation network, it will either be involved in an accident or it will not.

Thus, $X \sim B(1,p)$, where p is the probability of an accident when a vehicle enters any transportation network. In general, if n vehicles are passing through the transportation network (n trials) they are considered records of NOA in n trials, thus, $X \sim B(n,p)$. However, it was observed that the chance that a typical vehicle will cause an accident is very small out when considering the millions of vehicles that enter a transportation network (large number of n trials). Therefore, a $B(n,p)$ model for X is approximated by a $P(\lambda)$ model, where λ represents expected number of accidents. This approximation works well when λs are constant, but it is not reasonable to assume that λ across drivers and road segments are constant; in reality, this varies with each driver-vehicle combination. Considering NOA from different roads with different probabilities of accidents for drivers, the distribution of accidents have often been observed over-dispersed: if this occurs, $P(\lambda)$ is unlikely to show a good fit. In these cases, the negative binomial model can improve statistical fit to the process. In the literature, it has been suggested that over-dispersed processes may also be characterized by excess zeroes (more zeroes than expected under the $P(\lambda)$ process) and zero-inflated models can be a statistical solution to fit these types of processes.

Traffic Accident Data

The first data set analyzed was the number of accidents (NOA) causing death in the Dhaka district per month; NOA were counted for each of 64 months for the period of January 2003 to April 2008 and the data are presented in Table 4.1.1 and in Figure 4.1.1.

Table 4.1.1: Probability Distribution of NOA

NOA	Observed Months
0	0.12
1	0.24
2	0.17
3	0.22
4	0.17
5	0.05
6	0
7	0.02
Total	64 months

Table 4.1: Dataset Properties

Type	Time Period	n	Data Source
The number of traffic accidents (NOA) in the Dhaka district per month	Jan-2003 to April-2008	64 months	The Daily Star Newspaper
The number of peoples visiting (NOPV) Dhaka BMSSU per day	April-2007 to July-2007	74 days	BMSSU, Dhaka.
The number of earthquakes (NEQ) in Bangladesh per year	1973 to 2008	37 years	http://neic.usgs.gov/cgi-bin/epic/epic.cgi
The number of hartals (NOH) in the city of Dhaka per month	Jan-1972 to Dec-2007	432 months	Dasgupta (2001) and the Daily Star Newspaper

A total of 141 accidents occurred during the considered periods (see Figure 4.1.1) and the rate of accidents per month was 2.2 (see Table 4.1.2). Figure 4.1.1 also shows that since 2004 the rates have decreased. Main causes of road accidents identified according to Haque, 2003 include: rapid increase in the number of vehicles, more paved roads leading to higher speeds, poor driving and road use knowledge, skill and awareness and poor traffic management. The observed and expected frequencies with GOF statistic values are tabulated and shown in Table 4.1.2. The sample mean and variance indicate that the data shows over dispersion; about 16% of zeroes are present in the NOA data set. According to the GOF statistic, the binomial model fits poorly, whereas the Poisson and the negative binomial appear to fit. The excellent fits of different models are illustrated in Figure 4.1.2; based on the figure and the GOF statistic, the negative binomial model was shown to best fit NOA.

Figure 4.1.1: NOA by Year

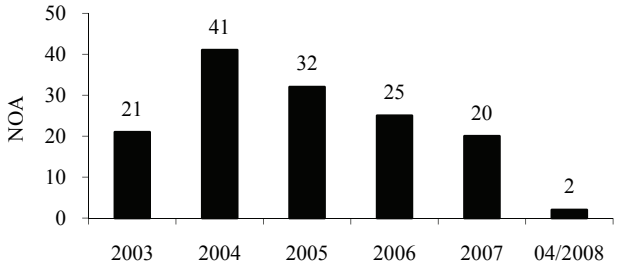


Table 4.1.2: Observed and Fitted Frequencies of NOA

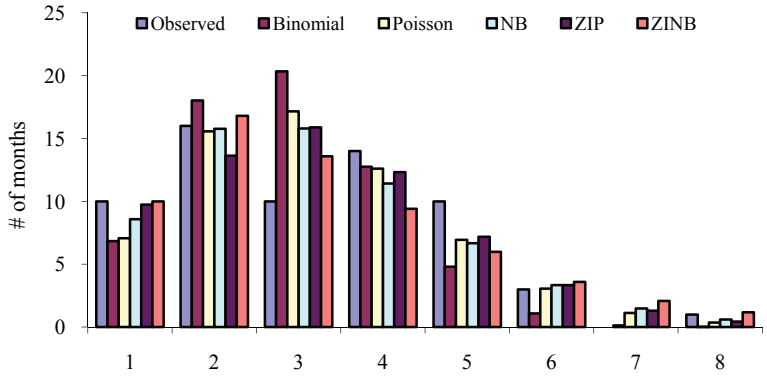
NOA	0	1	2	3	4	5	6	7	Chi-Square GOF Statistic
Observed Months	10	16	10	14	10	3	0	1	
B_{mom}	6.8	18.2	20.3	12.7	4.8	1.0	0.1	0	150.93*
B_{ml}	4.5	14.6	20.1	15.3	7.0	1.9	0.2	0	63.08*
Poisson	7.0	15.5	17.1	12.5	6.9	3.0	1.1	0.3	8.02
NB_{mom}	8.4	15.7	15.8	11.5	6.9	3.3	1.4	0.5	6.43
NB_{ml}	8.5	15.7	15.8	11.4	6.6	3.3	1.4	0.6	6.39
ZIP_{mom}	9.7	13.6	15.8	12.3	7.1	3.3	1.3	0.4	6.01
ZIP_{ml}	10.0	16.5	16.8	11.4	5.8	2.3	0.8	0.2	9.90
$ZINB_{ml}$	10.0	16.8	13.5	9.4	6.0	3.6	2.0	1.1	8.11

Mean = 2.2 and Variance = 2.6

Parameter Estimates: $B_{\hat{p}(mom)} = 0.27$, $B_{\hat{p}(ml)} = 0.31$, $P_{\hat{\lambda}(mom/ml)} = 2.2$, $NB_{\hat{p}(mom)} = 0.84$, $NB_{\hat{k}(ml)} = 11.96$, $NB_{\hat{p}(mom)} = 0.83$, $NB_{\hat{k}(ml)} = 11.09$, $ZIP_{\hat{p}(mom)} = 0.84$, $ZIP_{\hat{\lambda}(mom)} = 2.32$, $ZIP_{\hat{p}(ml)} = 0.84$, $ZIP_{\hat{\lambda}(ml)} = 2.03$, $ZINB_{\hat{\theta}(ml)} = 0.84$, $ZINB_{\hat{p}(ml)} = 0.53$, $ZINB_{\hat{k}(ml)} = 2.49$

Note: See footnotes, Table 3.1

Figure 4.1.2: Distribution of NOA for Different Models



Number of Patient Visits (NOPV) at Hospital

The number of patient visits (NOPV) data were collected from the medical unit of the Dhaka BMSSU medical hospital for the period of 26 April 2007 to 23 July 2007, where the variable of interest is the total number of patients visit in BMSSU per day. The frequency distribution for NOPV is reported in Table 4.2.1, which shows that the patients visiting rate per day is 142.36; this equates to a rate of 14.23 per working hour (see Table 4.2.2). Expected frequencies and GOF statistic values were tabulated and are shown in Table 4.2.2 and Figure 4.2.2 shows a bar chart of observed vs. expected frequencies. Tabulated results and the chart show that the negative binomial model and the ZTNB model (ZTNB model had the best fit) fit NOPV data well compared to other models. Based on this analysis, the ZTNB model is recommended to accurately fit NOPV per day.

Table 4.2.1: Frequency Distribution of NOPV

NOPV	Observed Days
51-83	1
84-116	12
117-149	32
150-182	23
183-215	6

Figure 4.2.1: Trend to Visits in BMSSU per Day

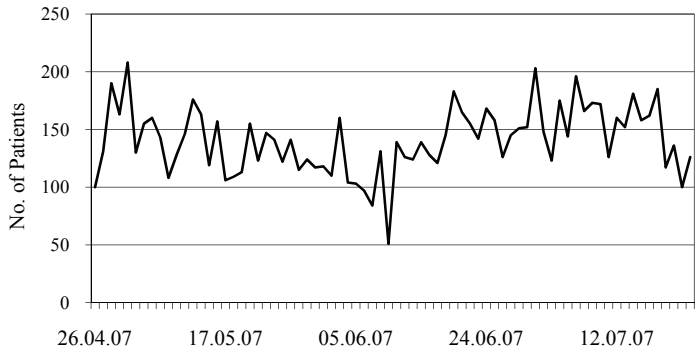


Table 4.2.2: Observed and Fitted Frequencies of NOPV

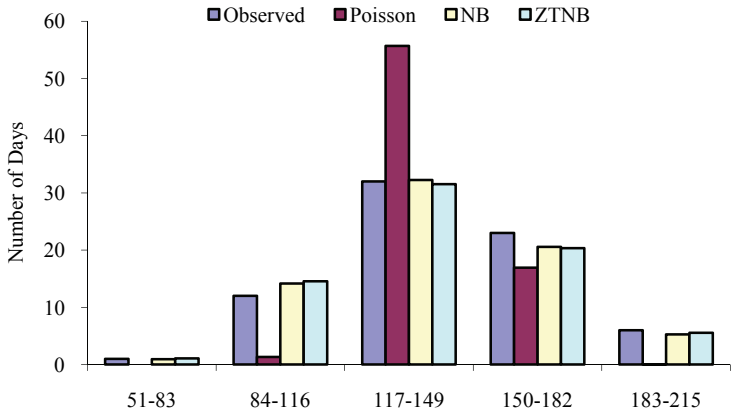
NOPV	51-83	84-116	117-149	150-182	183-215	Chi-Square GOF Statistic
Observed Days	1	12	32	23	6	
B_{mom}	0	0.02	66.22	7.75	0	1.5e+013*
B_{ml}	0	0.02	66.22	7.75	0	1.5e+013*
Poisson	0	1.33	55.69	16.94	0.02	1.46e+005*
NB_{mom}	0.92	14.17	32.27	20.58	5.28	0.7196
NB_{ml}	1.08	14.56	31.54	20.35	5.55	0.8453
ZIP_{mom}	73.03	0	0	0	0	-
ZIP_{ml}	0	1.33	55.69	16.94	0.02	1.46e+005*
$ZINB_{ml}$	0	0	0	0	0	-
ZTP_{ml}	0	1.33	55.69	16.94	0.02	1.46e+005
$ZTNB_{ml}$	1.08	14.56	31.54	20.35	5.55	0.84

Mean =142.36 and Variance =831.87

Parameter estimates: $B_{\hat{p}(mom)} = 0.65$, $B_{\hat{P}(ml)} = 0.65$, $P_{\hat{\lambda}(mom/ml)} = 140.78$, $NB_{\hat{p}(mom)} = 0.16$, $NB_{\hat{k}(mom)} = 26.82$, $NB_{\hat{p}(ml)} = 0.16$, $NB_{\hat{k}(ml)} = 26.82$, $ZIP_{\hat{p}(mom)} = 0.01$, $ZIP_{\hat{\lambda}(mom)} = 1.08e+004$, $ZIP_{\hat{p}(ml)} = 1.0$, $ZIP_{\hat{\lambda}(ml)} = 0.14$, $ZINB_{\hat{\theta}(ml)} = 1.0$, $ZINB_{\hat{p}(ml)} = 0.14$, $ZINB_{\hat{k}(ml)} = 831.87$, $ZTP_{\hat{\lambda}(ml)} = 140.78$, $ZTNB_{\hat{p}(ml)} = 0.14$ and $ZTNB_{\hat{k}(ml)} = 831.87$

Note: See footnotes, Table 3.1

Figure 4.2.2: Distribution of NOPV for Different Models



Earthquake Data

The third variable of interest is the number of earthquakes (NEQ) that occurred in Bangladesh from 1973 to January 2008 (based on available data). This data set was extracted from the http://earthquake.usgs.gov site and is presented in Table 4.3.1. The number of earthquakes per year is presented in Figure 4.3.1 and their magnitudes are displayed in Figure 4.3.2. The frequency distribution of earthquakes in Bangladesh is shown in Table 4.3.1. Table 4.3.2 shows a total of 127 earthquakes occurred in Bangladesh during the selected time period and that the average yearly earthquake rate is 3.43. The observed frequencies, the expected frequencies and the GOF statistic values for NEQ data are reported in Table 4.3.2. Sample mean and variance equal 3.43 and 10.19 respectively (shows over dispersion). It was found that the negative binomial model fits this data well (see Figure 4.3.3), whereas other models indicate lack of fit. Thus, based on this study, the distribution of NEQ follows the negative binomial distribution with a proportion of earthquakes equaling 0.29 per year.

Figure 4.3.1: Number of Earthquakes in Bangladesh per Year

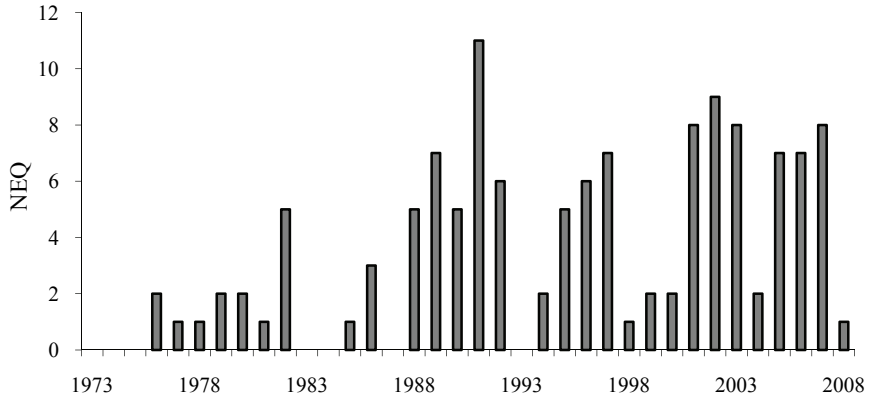


Figure 4.3.2: Earthquake Magnitudes

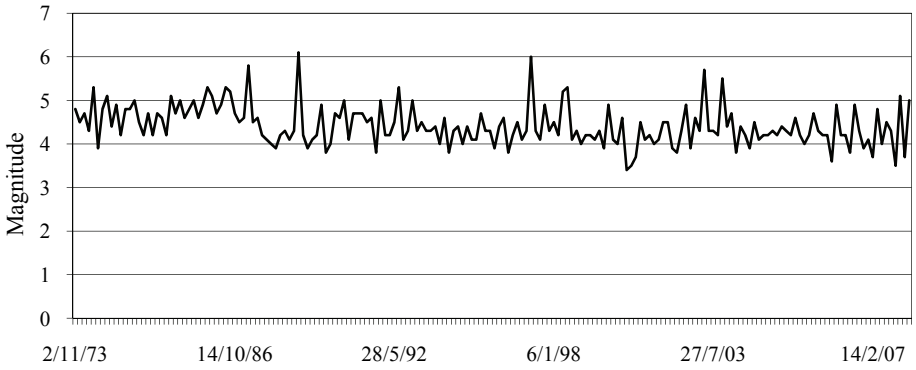


Table 4.3.1: Frequency Distribution of NEQ

NEQ	0	1	2	3	4	5	6	7 or More
Number of Years	7	6	7	1	0	4	2	9

Table 4.2.2: Observed and Fitted Frequencies of NOPV

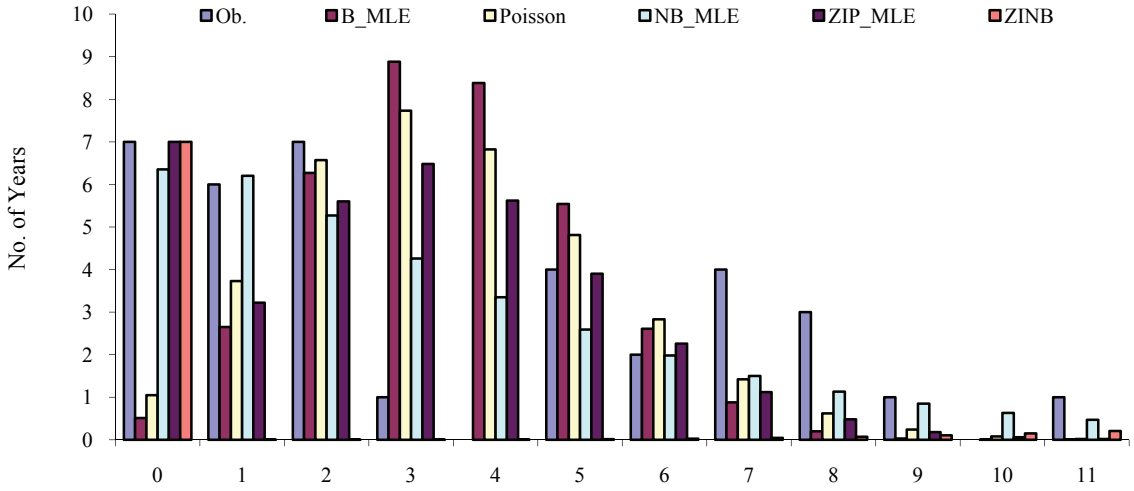
NEQ	Ob.	B_{mom}	B_{ml}	Poisson	NB_{mom}	NB_{ml}	ZIP_{ml}	ZIP_{ml}	$ZINB_{ml}$
0	7	0.78	0.51	1.05	4.96	6.35	12.32	7.00	7.00
1	6	3.58	2.65	3.73	6.26	6.20	0.60	3.22	0.0002
2	7	7.45	6.27	6.57	5.68	5.27	1.62	5.60	0.0009
3	1	9.31	8.88	7.73	4.78	4.26	2.89	6.48	0.0026
4	0	7.75	8.38	6.82	3.80	3.35	3.86	5.62	0.0065
5	4	4.52	5.54	4.81	2.92	2.59	4.12	3.90	0.0137
6	2	1.88	2.61	2.83	2.18	1.98	3.67	2.26	0.0258
7	4	0.56	0.88	1.42	1.60	1.50	2.80	1.12	0.0443
8	3	0.11	0.20	0.62	1.16	1.13	1.86	0.48	0.0708
9	1	0.01	0.03	0.24	0.83	0.85	1.10	0.18	0.1064
10	0	0	0	0.08	0.59	0.63	0.59	0.06	0.1518
11	1	0	0	0.02	0.41	0.47	0.28	0.02	0.2071
GOF Statistic		1.9e+004*	7.7e+003*	97.72*	16.23	15.25	77.09*	83.67*	2.28e+005

Mean =3.43 and Variance =10.19

Parameter estimates: $B_{\hat{p}(mom)} = 0.29$, $B_{\hat{p}(ml)} = 0.32$, $P_{\hat{\lambda}(mom/ml)} = 3.52$, $NB_{\hat{p}(mom)} = 0.34$, $NB_{\hat{k}(mom)} = 1.86$, $NB_{\hat{p}(ml)} = 0.27$, $NB_{\hat{k}(ml)} = 1.35$, $ZIP_{\hat{p}(mom)} = 0.65$, $ZIP_{\hat{\lambda}(mom)} = 5.33$, $ZIP_{\hat{p}(ml)} = 0.80$, $ZIP_{\hat{\lambda}(ml)} = 3.47$, $ZINB_{\hat{\theta}(ml)} = 0.80$, $ZINB_{\hat{p}(ml)} = 0.25$, $ZINB_{\hat{k}(ml)} = 10.19$.

Note: See footnotes, Table 3.1

Figure 4.3.3: Distribution of NEQ for Different Models



Hartal (Strike) Data

The fourth variable is the number of hartals (NOH) per month observed in Dhaka city from 1972 to 2007. Data from 1972 to 2000 was collected from Dasgupta (2001) and from 2001-2007 was collected from the daily newspaper, the Daily Star. Historically, the hartal phenomenon has respectable roots in Ghandi's civil disobedience against British colonialism (the word hartal, derived from Gujarati, is closing down shops or locking doors). In Bangladesh today, hartals are usually associated with the stoppage of vehicular traffic, closure of markets, shops, educational institutions and offices for a specific period of time to articulate agitation (Huq, 1992). When collecting monthly NOH data, care was taken to include all events that were consistent with the above definition of hartal (e.g., a hartal lasting 4 to 8 hours was treated as a half-day hartal, 9 to 12 hours as a full-day hartal; for longer hartals, each 12 hour period was treated as a full-day hartal). Historical patterns of hartals in Dhaka city, NOH with respect to time are plotted in Figure 4.4.1, and the frequency distribution of NOH is shown in Table 4.4.1. Between 1972 and 2007, 413 hartals were observed and the monthly hartal rate is 0.96 per month (see Table 4.4.2). Figure 4.4.1 shows the NOH for two periods: 1972-1990 (post-independence) and 1991-2007 (parliamentary democracy). It has been observed that the NOH have not decreased since the Independence in 1971. Although there were relatively few hartals in the early years following independence, the NOH began to rise sharply after 1981, with 101 hartals between 1982 and 1990. Since 1991(during the parliamentary democracy), the NOH have continued to rise with 125 hartals occurring from 1991-1996. Thus, the democratic periods (1991-1996 and 2003-2007) have experienced by far the largest number of hartals. Lack of political stability was found to be the main cause for this higher frequency of hartals (for details, see Beyond Hartals, 2005, p. 11). From Table 4.4.1, it may be observed that the hartal data contains about 60% of zeroes. Table 4.4.2 indicates that NOH process displays over-dispersion with a variance to mean > 1. According to data in this study (Table 4.4.2), the negative binomial distribution to model NOH with 31% chance of hartal per month is recommended.

Figure 4.4.1: Total Hartals in Dhaka City: 1972-2007

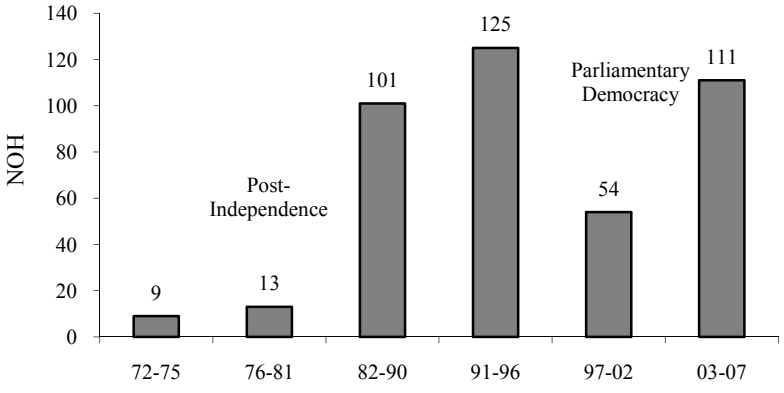


Table 4.4.1: Frequency Distribution of NOH

NOH	Number of Months
0	257
1	86
2	41
3	14
4	9
5	8
6	10
7 or More	7

Table 4.4.2: Observed and Fitted Frequencies of NOH

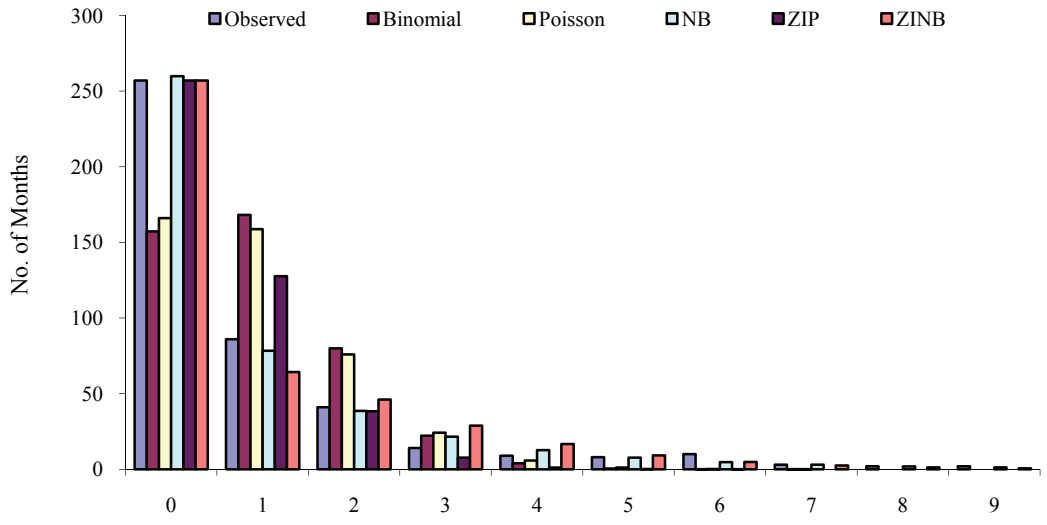
NOH	0	1	2	3	4	5	6	7	8	9	Chi-Square GOF Statistic
Observed Months	257	86	41	14	9	8	10	3	2	2	
B_{mom}	179.76	165.54	67.75	16.17	2.48	0.25	0.01	0.008	0	0	1.8e+007*
B_{ml}	157.23	168.18	79.95	22.17	3.95	0.47	0.03	0.001	0	0	5.4e+006*
Poisson	166.06	158.76	75.89	24.18	5.78	1.10	0.17	0.02	0	0	1.5e+004*
NB_{mom}	267.77	72.94	36.01	20.43	12.35	7.73	4.96	3.23	2.13	1.42	11.78
NB_{ml}	259.92	78.32	38.61	21.50	12.66	7.70	4.78	3.01	1.91	1.23	10.79
ZIP_{mom}	296.82	24.80	34.51	32.01	22.27	12.39	5.75	2.28	0.79	0.24	194.87*
ZIP_{ml}	257	127.66	38.34	7.67	1.15	0.13	0.01	0.00	0.00	0.00	7.3e+005*
$ZINB_{ml}$	257.00	64.28	46.11	28.80	16.65	9.17	4.87	2.53	1.28	0.64	27.88*

Mean =0.96 and Variance =3.35

Parameter estimates: $B_{\hat{P}(mom)} = 0.09$, $B_{\hat{P}(ml)} = 0.10$, $P_{\hat{\lambda}(mom/ml)} = 0.95$, $NB_{\hat{p}(mom)} = 0.28$, $NB_{\hat{k}(mom)} = 0.38$, $NB_{\hat{p}(ml)} = 0.31$, $NB_{\hat{k}(ml)} = 0.44$, $ZIP_{\hat{p}(mom)} = 0.31$, $ZIP_{\hat{\lambda}(mom)} = 2.78$, $ZIP_{\hat{p}(ml)} = 0.40$, $ZIP_{\hat{\lambda}(ml)} = 0.60$, $ZINB_{\hat{\theta}(ml)} = 0.40$, $ZINB_{\hat{p}(ml)} = 0.56$ and $ZINB_{\hat{k}(ml)} = 2.26$

Note: See footnotes, Table 3.1

Figure 4.4.2: Distribution of NOH for Different Models



Conclusion

This study reviewed some discrete models and compared them by assuming different amounts of zeroes in a sample. The following models were considered: the binomial model, the Poisson model, the negative binomial model and the zero-inflated and truncated models. A simulation study was conducted to observe the effects of excess zeroes on selected models, where data was generated from the Poisson model. This simulation study indicated that both the negative binomial and the ZIP models were useful to model discrete data with excess zeroes in the sample. Other models fit data containing excess zeroes poorly. Real-life examples were also used to illustrate the performance of the proposed models. All processes exhibited over-dispersion characteristic and could be fit well by the negative binomial model, with the exception of number of patients per day visiting a medical hospital, this data was better fit by ZTNB.

Acknowledgements

The authors are grateful to the management of the BMSSU medical college and hospital for providing records on the number of outpatient visits per day for the period of 26 April 2007 to 23 July 2007. We gratefully acknowledge contributions of the USGS staff for providing us with information about earthquakes in Bangladesh. Further we wish to thank the library staff of the Press Institute of Bangladesh for providing us information about hartal statistics. The article was partially completed while the second author was visited ISRT, Dhaka University, Bangladesh and ISI, Calcutta, India during July 2006.

References

Bliss, C. I., & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, *9*, 176-200.

Bohning, D. (1998). Zero-inflated Poisson models and C.A.M.A.N: A tutorial collection of evidence. *Biometrical Journal*, *40*, 833-843.

Dasgupta, A. (2001). *Sangbadpatrey Hartalchitra*. Press Institute of Bangladesh, Dhaka.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series* A, *222*, 309-368.

Fisher, R. A. (1941). The negative binomial distribution. *Annals of Eugenics*, *11*, 182-187.

Hoque, M. M. (2003). *Injuries from road traffic accidents: A serious health threat to the children*. Proceedings published on the World Health Day, 2003.

Huq, E. (1992). *Bangla academy byabaharik bangla abhidhan*. Dhaka: Bangla Academy.

Kibria, B. M. G. (2006). Applications of some discrete regression models for count data. *Pakistan Journal of Statistics and Operation Research*, *2*(*1*), 1-16.

Lloyd-Smith, J. D. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly over-dispersed data with applications to infectious diseases. *pLos one*, *2*, 1-8.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*, 35-46.

MATLAB (version 7.0), MathWorks, New York, 2004-2007.

Neyman, J. (1939). On a new class of contagious distribution, applicable in entomology and bacteriology. *Annals of Mathematical Statistics*, *10*, 35-57.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, *Series A*, 71-110.

Rider, P. R. (1961). Estimating the parameters of mixed Poisson, binomial and Weibul distributions by method of moments. *Bulletin de l'Institut International de Statistiques*, *38*, Part 2.

Ross, G. J. S., & Preece, D. A. (1985). The negative binomial distribution. *The Statistician*, *34*, 323-336.

Shankar, V. N., Ulfarsson, G. F., Pendyala, R. M., & Nebergal, M. B. (2003). Modeling crashes involving pedestrians and motorized traffic. *Safety Science*, *41*, 627-640.

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, *189*, 732-735.

United Nations. (2005). *Beyond Hartals: Towards Democratic Dialogue in Bangladesh*. United Nations Development Program Bangladesh, ISBN 984-32-1424-2, March 2005.

Warton, D. I. (2005). Many zeroes do not mean zero inflation: comparing the goodness of-fit of parametric models to multivariate abundance data. *Environmetris*, *16*, 275-289.

White, G. C., & Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology*, *77*, 2549-2557.

Zhang, Y. Z., & Lord, D. (2007). Estimating dispersion parameter of negative binomial distribution for analyzing crash data using bootstrapped maximum likelihood method, *Journal of Transportation Research Board,* Issue Number: *2019*, 15-21.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 448-462

1538 – 9472/09/$95.00

Impact of Rank-Based Normalizing Transformations on the Accuracy of Test Scores

Shira R. Solomon
CNA Education

Shlomo S. Sawilowsky
Wayne State University

The purpose of this article is to provide an empirical comparison of rank-based normalization methods for standardized test scores. A series of Monte Carlo simulations were performed to compare the Blom, Tukey, Van der Waerden and Rankit approximations in terms of achieving the T score's specified mean and standard deviation and unit normal skewness and kurtosis. All four normalization methods were accurate on the mean but were variably inaccurate on the standard deviation. Overall, deviation from the target moments was pronounced for the even moments but slight for the odd moments. Rankit emerged as the most accurate method among all sample sizes and distributions, thus it should be the default selection for score normalization in the social and behavioral sciences. However, small samples and skewed distributions degrade the performance of all methods, and practitioners should take these conditions into account when making decisions based on standardized test scores.

Key words: Normalization; normalizing transformations; T scores; test scoring; ranking methods; Rankit; Blom; Tukey; Van der Waerden; Monte Carlo.

Introduction

Standardization and normalization are two ways of defining the frame of reference for a distribution of test scores. Both types of score conversions, or transformations, mathematically modify raw score values (Osborne, 2002). The defining feature of standard scores is that they use standard deviations to describe scores' distance from the mean, thereby creating equal units of measure within a given score distribution. Standard scores may be modified to change the scale's number system (Angoff, 1984), but unless distributions of standard scores are normalized, they will retain the shape of the original score distribution. Therefore, standardization may enable effective analysis of individual scores within a single test, but normalization is needed for meaningful comparisons between tests.

Shira R. Solomon is a Research Analyst. Email: solomons@cna.org. Shlomo S. Sawilowsky is Professor of Evaluation and Research. Email: shlomo@wayne.edu.

The Problem of Non-continuous Data in Educational and Psychological Testing

Knowledge, intellectual ability, and personality are psychological objects that can only be measured indirectly, not by direct observation (Dunn-Rankin, 1983). The scales that describe them are hierarchical—they result in higher or lower scores—but these scores do not express exact quantities of test-takers' proficiency or attitudes. Ordinal test items such as Likert scales result in raw scores that are meaningless without purposeful statistical interpretation (Nanna & Sawilowsky, 1998). Measures with unevenly spaced increments interfere with the interpretation of test scores against performance benchmarks, the longitudinal linking of test editions, and the equating of parallel forms of large-scale tests (Aiken, 1987). They also threaten the robustness and power of the parametric statistical procedures that are conventionally used to analyze standardized test scores (Friedman, 1937; Sawilowsky & Blair, 1992).

Statisticians have been transforming ordinal data into a continuous scale since Fisher and Yates tabled the normal deviates in 1938. Wimberly (1975) favored rank-based

transformations to other normalizing transformations such as those based on logarithms, exponents, or roots for their superior accuracy among random scores of different variables. Rank-based transformations not only attempt to equate the means and homogenize the variance of test score distributions, they also aim to create conformity in the third and fourth moments, skewness and kurtosis. Central tendency and variability have clear implications for test score distributions.

The most prominent of the rank-based normalization procedures, based on their inclusion in widely used statistical software (e.g., SPSS, 2006) are those attributed to Van der Waerden, Blom, Bliss (the Rankit procedure), and Tukey. Van der Waerden's formula (1952, 1953a, 1953b; Lehmann, 1975) was thought to improve on percentiles by computing quantiles (equal unit portions under the normal curve corresponding with the number of observations in a sample) not strictly on the basis of ranks, but according to the rank of a given score value relative to the sample size (Conover, 1980). Blom's formula (1958) responds to the curvilinear relationship between a score's rank in a sample and its normal deviate. Because "Blom conjectured that α always lies in the interval (0·33, 0·50)," explained Harter, "he suggested the use of α = 3/8 as a compromise value" (1961, p.154). Bliss, Greenwood, and White (1956) credited Ipsen and Jerne (1944) with coining the term "rankit," but Bliss is credited with developing the technique as it is now used. Bliss, et al. refined this approximation in their study of the effects of different insecticides and fungicides on the flavor of apples. Its design drew on Scheffé's advancements in paired comparison research, which sought to account for magnitude and direction of preference, in addition to preference itself. Tukey may have proposed his formula, which he characterized as "simple and surely an adequate approximation to what is claimed to be optimum" (1962, p.22), as a refinement of Blom's.

These procedures have been explored to various degrees in the context of hypothesis testing, where the focus is necessarily on their properties in the tails of a distribution. In the context of standardized testing, however, the body of the distribution—that is, the 95% of the curve that lies between the tails—is the focus. Practitioners need to know how accurately each method normalizes non-theoretical score distributions. Solomon (2008) produced the first empirical comparison of the Van der Waerden, Blom, Tukey, and Rankit methods as they apply to standardized testing. This study sought to demonstrate their accuracy under a variety of sample size and distributional conditions.

Table 1: Chronology of Rank-Based Normalization Procedure Development

Procedure	Year	Formula
Van der Waerden	1952	r* / (n + 1)
Blom	1954	(r - 3/8) / (n + 1/4)
Rankit	1956	(r - 1/2) / n
Tukey	1962	(r - 1/3) / (n + 1/3)

*where r is the rank, ranging from 1 to n

Blom, Tukey, Van der Waerden, and Rankit each contribute a formula that approximates a normal distribution, given a set of raw scores or non-normalized standard scores. However, the formulas themselves had not been systematically compared for their first four moments' accuracy in terms of normally distributed data. Nor had they been compared in the harsher glare of non-normal distributions, which are prevalent in the fields of education and psychology (Micceri, 1989). Small samples are also common in real data and are known to have different statistical properties than large samples (Conover, 1980). In general, real data can be assumed to behave differently than data that is based on theoretical distributions, even if these are non-normal (Stigler, 1977).

A series of Monte Carlo simulations drew samples of different sizes from eight unique, empirically established population distributions. These eight distributions, though extensive in their representation of real achievement and psychometric test scores, do not represent all possible distributions that could occur in educational and psychological testing or in social and behavioral science investigations

more generally. Nor do the sample sizes represent every possible increment. However, both the sample size increments and the range of distributional types are assumed to be sufficient for the purpose of outlining the absolute and comparative accuracy of these normalizing transformations in real settings. Although the interpretation of results need not be restricted to educational and psychological data, similar distributional types may be most often found in these domains.

For normally distributed variables, the standardization process begins with the *Z* score transformation, which produces a mean of 0 and a standard deviation of 1 (Walker & Lev, 1969; Mehrens & Lehmann, 1980; Hinkle, Wiersma, & Jurs, 2003). *Z* scores are produced by dividing the deviation score (the difference between raw scores and the mean of their distribution) by the standard deviation: $Z = (X - \mu) / \sigma$. However, *Z* scores can be difficult to interpret due to decimals and negative numbers. Because 95% of the scores fall between -3 and +3, small changes in decimals may imply large changes in performance. Also, because half the scores are negative, it may appear to the uninitiated that half of the examinees obtained an extremely poor outcome.

Linear versus Area Transformations

Linear scaling remedies these problems by multiplying standard scores by a number large enough to render decimal places trivial, then adding a number large enough to eliminate negative numbers. Although standard scores may be assigned any mean and standard deviation through linear scaling, the *T* score scale ($S_T = 10Z + 50$) has dominated the scoring systems of social and behavioral science tests for a century (Cronbach, 1976; Kline, 2000; McCall, 1939). In the case of a normally distributed variable, the resulting *T*-scaled standard scores would have a mean of 50 and a standard deviation of 10. In the context of standardized testing, however, *T* scores refer not to *T*-scaled standard scores but to *T*-scaled normal scores. In the *T* score formula, *Z* refers to a score's location on a unit normal distribution—its normal deviate—not its place within the testing population.

Scaling standard scores of achievement and psychometric tests has limited value. Most educational and psychological measurements are ordinal (Lester & Bishop, 2000), but standard scores can only be obtained for continuous data because they require computation of the mean. Furthermore, linear transformations retain the shape of the original distribution. If a variable's original distribution is Gaussian, its transformed distribution will also be normal. If an observed distribution manifests substantial skew, excessive or too little kurtosis, or multimodality, these non-Gaussian features will be maintained in the transformed distribution.

This is problematic for a wide range of practitioners because it is common practice for educators to compare or combine scores on separate tests and for testing companies to reference new versions of their tests to earlier versions. Standard scores such as *Z* will not suffice for these purposes because they do not account for differing score distributions between tests. Comparing scores from a symmetric distribution with those from a negatively skewed distribution, for example, will give more weight to the scores at the lower range of the skewed curve than to those at the lower range of the symmetric curve (Horst, 1931). Normalizing transformations are used to avoid biasing test score interpretation due to heteroscedastic and asymmetrical score distributions.

Non-normality Observed

According to Nunnally (1978), "test scores are seldom normally distributed" (p.160). Micceri (1989) demonstrated the extent of this phenomenon in the social and behavioral sciences by evaluating the distributional characteristics of 440 real data sets collected from the fields of education and psychology. Standardized scores from national, statewide, and districtwide test scores accounted for 40% of them. Sources included the Comprehensive Test of Basic Skills (CTBS), the California Achievement Tests, the Comprehensive Assessment Program, the Stanford Reading tests, the Scholastic Aptitude Tests (SATs), the College Board subject area tests, the American College Tests (ACTs), the Graduate Record Examinations (GREs), Florida Teacher Certification Examinations for adults, and

Florida State Assessment Program test scores for 3rd, 5th, 8th, 10th, and 11th grades.

Micceri summarized the tail weights, asymmetry, modality, and digit preferences for the ability measures, psychometric measures, criterion/mastery measures, and gain scores. Over the 440 data sets, Micceri found that only 19 (4.3%) approximated the normal distribution. No achievement measure's scores exhibited symmetry, smoothness, unimodality, or tail weights that were similar to the Gaussian distribution. Underscoring the conclusion that normality is virtually nonexistent in educational and psychological data, none of the 440 data sets passed the Kolmogorov-Smirnov test of normality at alpha = .01, including the 19 that were relatively symmetric with light tails. The data collected from this study highlight the prevalence of non-normality in real social and behavioral science data sets.

Furthermore, it is unlikely that the central limit theorem will rehabilitate the demonstrated prevalence of non-normal data sets in applied settings. Although sample means may increasingly approximate the normal distribution as sample sizes increase (Student, 1908), it is wrong to assume that the original population of scores is normally distributed. According to Friedman (1937), "this is especially apt to be the case with social and economic data, where the normal distribution is likely to be the exception rather than the rule" (p.675).

There has been considerable empirical evidence that raw and standardized test scores are non-normally distributed in the social and behavioral sciences. In addition to Micceri (1989), numerous authors have raised concerns regarding the assumption of normally distributed data (Pearson, 1895; Wilson & Hilferty, 1929; Allport, 1934; Simon, 1955; Tukey & McLaughlin, 1963; Andrews et al., 1972; Pearson & Please, 1975; Stigler, 1977; Bradley, 1978; Tapia & Thompson, 1978; Tan, 1982; Sawilowsky & Blair, 1992). The prevalence of non-normal distributions in education, psychology, and related disciplines calls for a closer look at transformation procedures in the domain of achievement and psychometric test scoring.

The Importance of T Scores for the Interpretation of Standardized Tests

Standardized test scores are notoriously difficult to interpret (Chang, 2006; Kolen and Brennan, 2004; Micceri, 1990; Petersen, Kolen, and Hoover, 1989). Most test-takers, parents, and even many educators, would be at a loss to explain exactly what a score of 39, 73, or 428 means in conventional terms, such as pass/fail, percentage of questions answered correctly, or performance relative to other test-takers. Despite the opaqueness of *T* scores relative to these conventional criteria, they have the advantage of being the most familiar normal score scale, thus facilitating score interpretation. Most normal score systems are assigned means and standard deviations that correspond with the *T* score. For example, the College Entrance Board's *Scholastic Aptitude Test* (SAT) Verbal and Mathematical sections are scaled to a mean of 500 and a standard deviation of 100. *T* scores fall between 20 and 80 and SAT scores fall between 200 and 800. The *T* score scale facilitates the interpretation of test scores from any number of different metrics, few of which would be familiar to a test taker, teacher, or administrator, and gives them a common framework.

The importance of transforming normal scores to a scale that preserves a mean of 50 and a standard deviation of 10 calls for an empirical comparison of normalizing transformations. This study experimentally demonstrates the relative accuracy of the Blom, Tukey, Van der Waerden, and Rankit approximations for the purpose of normalizing test scores. It compares their accuracy in terms of achieving the *T* score's specified mean and standard deviation and unit normal skewness and kurtosis, among small and large sample sizes in an array of real, non-normal distributions.

Methodology

A Fortran program was written to compute normal scores using the four rank-based normalization formulas under investigation. Fortran was chosen for its large processing capacity and speed of execution. This is important for Monte Carlo simulations, which typically require from thousands to millions of iterations.

Normal scores were computed for each successive iteration of randomly sampled raw scores drawn from various real data sets. The resulting normal scores were then scaled to the *T*. The first four moments of the distribution were calculated from these *T* scores for each of the 14 sample sizes in each of the eight populations. Absolute values were computed by subtracting *T* score means from 50, standard deviations from 10, skewness values from 0, and kurtosis values from 3. These absolute values were sorted into like bins and ranked in order of proximity to the target moments. The values and ranks were averaged over the results from 10,000 simulations and reported in complete tables (Solomon, 2008). Average root mean square (RMS) values and ranks were also computed and reported for the target moments. This paper summarizes the values and ranks for absolute deviation values and RMS, or magnitude of deviation. Together, deviation values and magnitude of deviation describe the accuracy and stability of the Blom, Tukey, Van der Waerden, and Rankit approximations in attaining the first four moments of the normal distribution.

Sample Sizes and Iterations

Simulations were conducted on samples of size n = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200, 500, and 1,000 that were randomly selected from each of the eight Micceri (1989) data sets. Ten-thousand (10,000) iterations were performed to break any ties up to three decimal places.

Achievement and Psychometric Distributions

Micceri (1989) computed three indices of symmetry/asymmetry and two indices of tail weight for each of the 440 large data sets he examined (for 70% of which, n $\geq$ 1,000), grouped by data type: achievement/ability (accounting for 231 of the measures), psychometric (125), criterion/mastery (35), and gain scores (49). Eight distributions were identified based on symmetry, tail weight contamination, propensity scores, and modality. Sawilowsky, Blair, and Micceri (1990) translated these results into a Fortran subroutine using achievement and psychometric measures that best represented the distributional characteristics described by Micceri (1989).

The following five distributions were drawn from achievement measures: Smooth Symmetric, Discrete Mass at Zero, Extreme Asymmetric – Growth, Digit Preference, and Multimodal Lumpy. Mass at Zero with Gap, Extreme Asymmetric – Decay, and Extreme Bimodal were drawn from psychometric measures. All eight achievement and psychometric distributions are nonnormal. These distributions are described in Table 2 and graphically depicted in Figure 1.

Results

The purpose of this study was to compare the accuracy of the Blom, Tukey, Van der Waerden, and Rankit approximations in attaining the target moments of the normal distribution. Tables 3, 4, and 5 present these results. Table 3 summarizes the major findings according to moment, sample size, and distribution. It presents values and simplified ranks for the accuracy of the four normalizing methods on the first measure, deviation from target moment. For example, the *T* score's target standard deviation is 10. Therefore, two methods that produce a standard deviation of 9.8 or 10.2 would have the same absolute deviation value: 0.2. The highest ranked method for each condition is the most accurate, having the smallest absolute deviation value over 10,000 Monte Carlo repetitions. It is possible to assign ranks on the mean despite the accuracy of all four normalization methods because differences begin to appear at six decimal places. However, all numbers are rounded to the third decimal place in the tables.

Table 3 shows that rank-based normalizing methods are less accurate on the standard deviation than on the mean, skewness, or kurtosis. Furthermore, the standard deviation has more immediate relevance to the interpretation of test scores than the higher moments. For these reasons, Tables 4 and 5 and Figures 2 and 3 restrict their focus to the methods' performance on the standard deviation. Table 4 summarizes the methods' proximity to the target standard deviation by distribution type. Table 5 reports proximity for all eight distributions.

Table 2: Basic Characteristics of Eight Non-normal Distributions

	Range	Mean	Median	Variance	Skewness	Kurtosis
	Achievement					
1. Smooth Symmetric	$0 \le x \le 27$	13.19	13.00	24.11	0.01	2.66
2. Discrete Mass at Zero	$0 \le x \le 27$	12.92	13.00	19.54	-0.03	3.31
3. Extreme Asymmetric – Growth	$4 \le x \le 30$	24.50	27.00	33.52	-1.33	4.11
4. Digit Preference	$420 \le x \le 635$	536.95	535.00	1416.77	-0.07	2.76
5. Multimodal Lumpy	$0 \le x \le 43$	21.15	18.00	141.61	0.19	1.80
	Psychometric					
	Range	Mean	Median	Variance	Skewness	Kurtosis
6. Mass at Zero w/Gap	$0 \le x \le 16$	1.85	0	14.44	1.65	3.98
7. Extreme Asymmetric – Decay	$10 \le x \le 30$	13.67	11.00	33.06	1.64	4.52
8. Extreme Bimodal	$0 \le x \le 5$	2.97	4.00	2.86	-0.80	1.30

Figure 1: Appearance of Five Achievement and Three Psychometric Distributions (Sawilowsky & Fahoome, 2003)

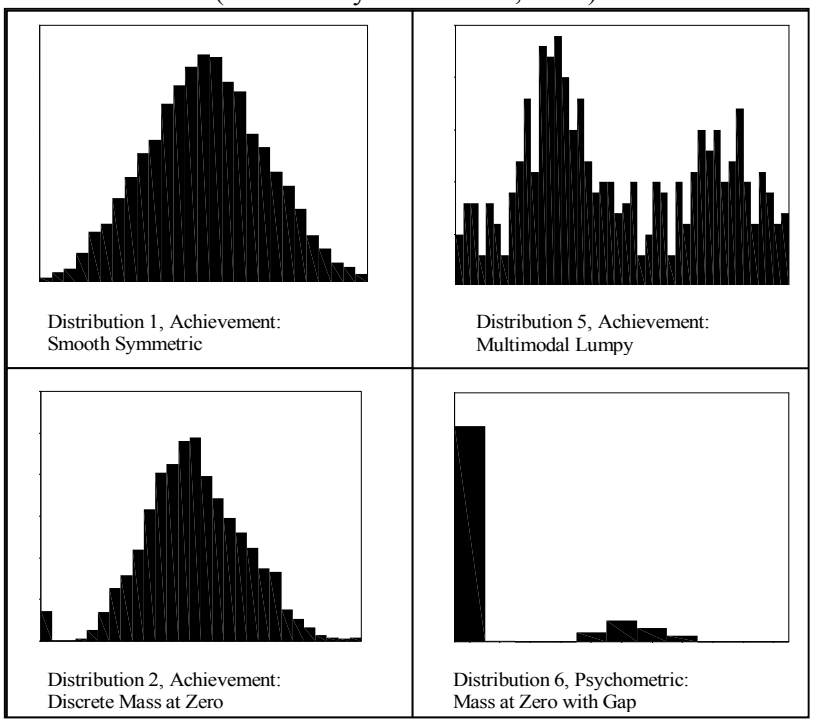


Figure 1 (Continued): Appearance of Five Achievement and Three Psychometric Distributions (Sawilowsky & Fahoome, 2003)

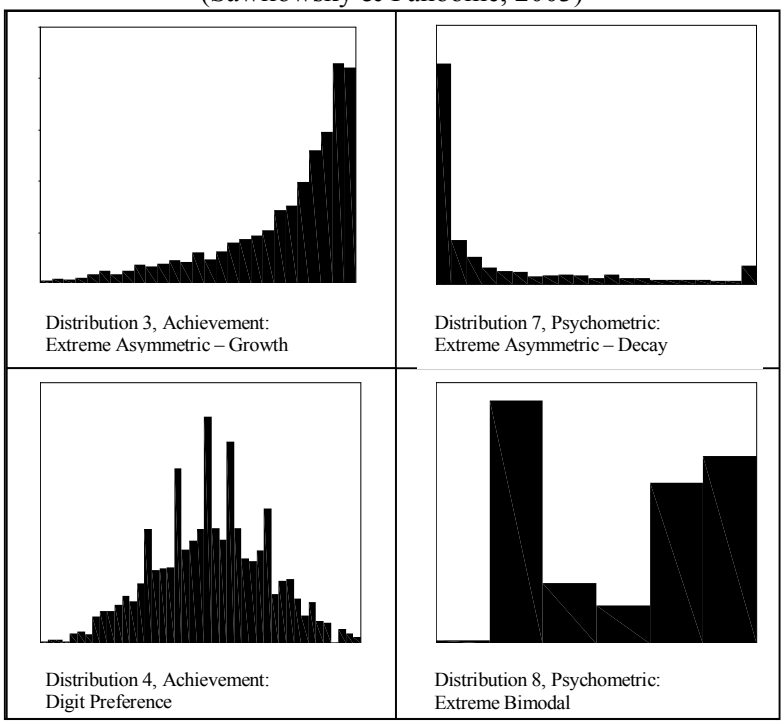


Table 3: Deviation from Target, Summarized by Moment, Sample Size and Distribution

	Moment							
	Blom		Tukey		Van der W.		Rankit	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Mean	4	0.000	1	0.000	2	0.000	3	0.000
Standard Dev	2	1.142	3	1.186	4	1.603	1	1.119
Skewness	2	0.192	2	0.192	1	0.191	2	0.192
Kurtosis	2	0.947	3	0.941	4	0.952	1	0.930
	Sample Size							
	Blom		Tukey		Van der W.		Rankit	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
$5 \leq 50$	2	0.609	3	0.628	4	0.769	1	0.603
$100 \leq 1000$	2	0.435	3	0.423	4	0.447	1	0.416
	Distribution							
	Blom		Tukey		Van der W.		Rankit	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Smooth Sym	2	0.393	3	0.411	4	0.531	1	0.391
Discr Mass Zero	2	0.404	3	0.421	4	0.539	1	0.403
Asym – Growth	2	0.453	3	0.470	4	0.583	1	0.452
Digit Preference	2	0.390	3	0.408	4	0.527	1	0.370
MM Lumpy	2	0.412	3	0.396	4	0.510	1	0.376
MZ w/Gap	2	1.129	3	1.126	4	1.204	1	1.113
Asym – Decay	2	0.726	3	0.739	4	0.835	1	0.725
Extr Bimodal	2	0.655	3	0.669	4	0.765	1	0.654

Proximity to target includes deviation values, at the top of the Tables 4 and 5, and RMS values, at the bottom. RMS is an important second measure of accuracy because it indicates how consistently the methods perform. By standardizing the linear distance of each observed moment from its target, RMS denotes within-method magnitude of deviation. Respectively, the two accuracy measures, deviation value and magnitude of deviation, describe each method's average distance from the target value and how much its performance varies over the course of 10,000 random events.

Predictive Patterns of the Deviation Range

Figure 2 plots the range of deviation values for each distribution against a power curve among small samples. Curve fitting is only possible for the deviation range on the second and fourth moments, standard deviation and kurtosis. The first and third moments, mean and skewness, either contain zeros, which make transformations impossible, or lack sufficient variability to make curve fitting worthwhile.

To evaluate trends at larger sample sizes, the small-sample regression models are fitted a second time with the addition of four sample sizes: n = 100, n = 200, n = 500, and n = 1000. To whatever extent predictive patterns are established when $n \leq 50$, those regression slopes either improve in fit or continue to hold when sample sizes increase. Figure 3 shows that inclusion of larger sample sizes causes the Smooth Symmetric power curve to remain intact and the Digit Preference power curve to improve in fit.

Table 4: Proximity to Target Standard Deviation for Achievement and Psychometric Distributions

Deviation Value

	Blom		Tukey		Van der Waerden		Rankit	
	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$
Achievement	0.736	0.205	0.824	0.122	1.413	0.231	0.735	0.089
Psychometric	2.263	1.382	2.332	1.390	2.802	1.455	2.260	1.374

Magnitude of Deviation (RMS)

	Blom		Tukey		Van der Waerden		Rankit	
	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$	$5 \leq 50$	$100 \leq 1000$
Achievement	0.018	0.001	0.017	0.001	0.017	0.001	0.009	0.001
Psychometric	0.542	0.096	0.540	0.096	0.536	0.096	0.497	0.088

Table 5: Proximity to Target Standard Deviation for Small and Large Samples

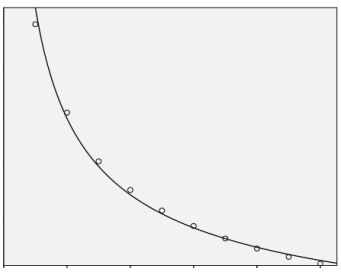
Deviation Value

	Blom		Tukey		Van der Waerden		Rankit	
	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000
Smooth Sym	0.720	0.077	0.808	0.089	1.401	0.202	0.719	0.047
Discr MZ	0.736	0.082	0.823	0.094	1.414	0.208	0.734	0.073
Asym – Gro	0.829	0.247	0.914	0.260	1.489	0.356	0.827	0.237
Digit Pref	0.702	0.072	0.790	0.084	1.385	0.195	0.700	0.043
MM Lumpy	0.696	0.547	0.785	0.085	1.378	0.196	0.695	0.044
MZ w/Gap	3.651	2.804	3.711	2.815	4.117	2.896	3.647	2.795
Asym – Dec	1.668	0.420	1.743	0.425	2.244	0.458	1.666	0.417
Extr Bimod	1.469	0.921	1.543	0.931	2.045	1.011	1.467	0.912

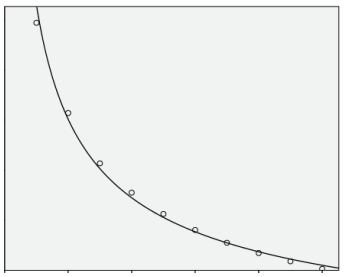
Magnitude of Deviation (RMS)

	Blom		Tukey		Van der Waerden		Rankit	
	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000	5 ≤ 50	100 ≤ 1000
Smooth Sym	0.003	0.000	0.003	0.000	0.003	0.000	0.003	0.000
Discr MZ	0.015	0.000	0.015	0.000	0.014	0.000	0.003	0.000
Asym – Gro	0.043	0.003	0.042	0.003	0.042	0.003	0.035	0.003
Digit Pref	0.013	0.000	0.014	0.000	0.013	0.000	0.003	0.000
MM Lumpy	0.013	0.000	0.013	0.000	0.013	0.000	0.002	0.000
MZ w/Gap	1.081	0.225	1.077	0.225	1.069	0.225	0.993	0.226
Asym – Dec	0.310	0.031	0.309	0.031	0.307	0.031	0.290	0.031
Extr Bimod	0.236	0.031	0.235	0.031	0.232	0.031	0.208	0.007

Figure 2: Power Curves Fitted to the Deviation Range of the Standard Deviation at 10 Small Sample Sizes

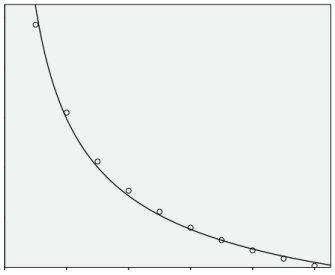


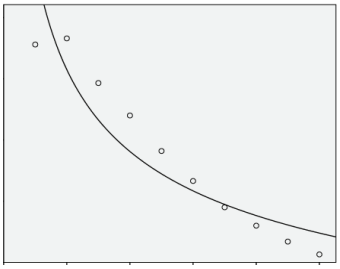
Smooth Symmetric

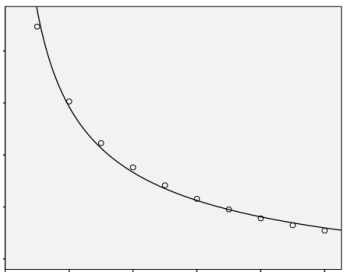


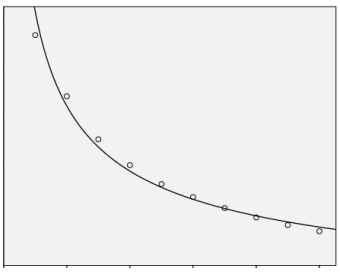
Multimodal Lumpy

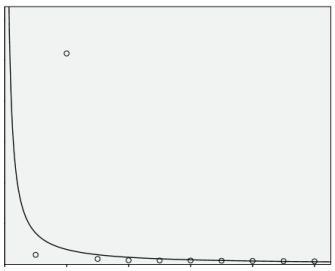
Figure 2 (Continued): Power Curves Fitted to the Deviation Range of the Standard Deviation at 10 Small Sample Sizes


Discrete Mass at Zero


Mass at Zero with Gap


Extreme Asymmetric Growth


Extreme Asymmetric Decay


Digit Preference

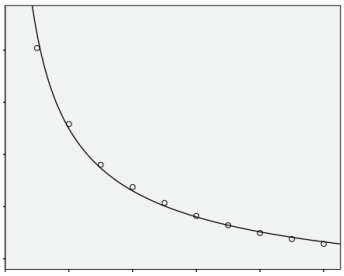
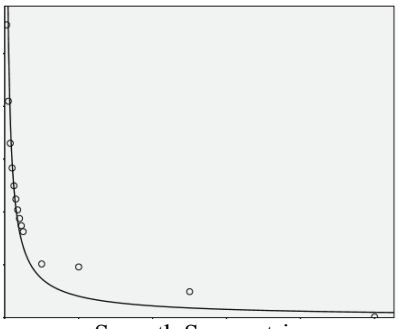
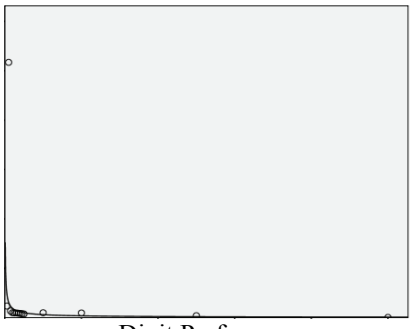

Extreme Bimodal

Figure 3: Power Curves Fitted to the Deviation Range of the Standard Deviation with Inclusion of Four Large Sample Sizes


Smooth Symmetric


Digit Preference

Conclusion

Table 3 shows that Rankit outperforms the other methods across moments at small and large sample sizes and with all eight distributions. Blom and Tukey consistently place second and third, and Van der Waerden performs the worst.

Mean, Skewness, and Kurtosis

All four rank-based normalization methods attain the target value of 50 for the mean. Differences appear in the numerical results only after the third decimal place, and are therefore meaningless in terms of practical application. These differences are reflected in the deviation ranks in Table 3. The four methods' average deviation from the target skewness value of the normal distribution is 0.192 (Table 3). Normalization methods should not be selected on the basis of their deviation from target skewness values because the deviation quantities are small and the differences between them are negligible.

Deviation values for kurtosis show greater deviation from target than those for skewness but less than those for standard deviation. The average deviation value for kurtosis across all sample sizes and distributions is 0.943 (Table 3). Moderate flatness or peakedness might reflect something about the test instrument or the population, but it is not clear how kurtosis could affect decisions made about test scores.

Standard Deviation: Deviation from Target Standard Deviation.

None of the Normalization methods attains the target standard deviation on either accuracy measure. Rankit is the most accurate method, averaging a distance of 1.119 from the target T score standard deviation of 10 (Table 3). This means that the practitioner who uses Rankit to normalize test scores without reference to sample size or distribution can expect to obtain an estimated standard deviation between 8.881 and 11.119. If Z = 2, the T score would fall between 67.762 or 72.238, for a range of 4.476. Adding in the test instrument's standard error compounds the problem. An instrument with a standard error of three (± 3) would expand the true score range by six points, to 10.476. Rounding to the nearest whole number, this means that the test-taker's standardized test score falls somewhere between 65 and 75. Even a standard error half this size would lead to a true score range of 7.476. Thus, a standard deviation that is off target by 1.119 would combine with a standard error of ± 1.5 to increase the true score range by 249%, from a theorized range of three to an actual range of seven and a half. As the standard error increases, the estimated difference between the theorized and actual score range diminishes. At a standard error of three, Rankit produces a standard deviation that causes the true score range to be 175% greater than the presumed score range.

Van der Waerden is the least accurate method, averaging a distance of 1.603 from the target *T* score standard deviation (Table 3). Using Van der Waerden to normalize a test score ($Z = 2$) without reference to sample size or distribution produces a rounded true score range of 64 to 76 at a standard error of ± 3. At a standard error of ± 1.5, the test-taker's *T* score would fall between 65 and 75, the same range that Rankit produced at twice the standard error. Van der Waerden's inaccuracy on the standard deviation causes the true score range to increase over the expected score range by 207% at a standard error of ± 3 and 314% at a standard error of ± 1.5.

As with Rankit, smaller standard errors correspond with greater relative inaccuracy of the true score range. The more reliable a test instrument is, the less precise are the *T* scores, regardless of the normalization method used. This is illustrated in Table 6, which presents the percentage increase to the true score range based on each method's overall distance from the standard deviation across all sample sizes and distributions.

The inaccuracy of the rank-based normalization methods on the standard deviation becomes more pronounced in the context of sample size and distribution type (Table 4). All four methods are more accurate among large samples and achievement distributions and less accurate among small samples and psychometric distributions. Rankit's worst average deviation value, among psychometric distributions at small sample sizes, is 25 times higher than its best among achievement distributions at large sample sizes.

Table 6: Increase of True Score Range over Expected Score Range by Standard Error

Standard Error	% Increase			
	Rankit	Blom	Tukey	Van der Waerden
± 0.5	548%	557%	574%	741%
± 1.0	324%	328%	337%	421%
± 1.5	249%	252%	258%	314%
± 2.0	212%	214%	219%	260%
± 2.5	190%	191%	195%	228%
± 3.0	175%	176%	179%	207%

Van der Waerden's worst deviation value — again, among psychometric distributions at small sample sizes — is 12 times higher than its best. Normalization performance is so heavily influenced by sample size and distribution type that Van der Waerden, which is the worst overall performer, produces much more accurate standard deviations under the best sample size and distributional conditions than Rankit does under the worst distributional conditions. Under these circumstances, Rankit's worst deviation value is 10 times higher than Van der Waerden's best deviation value.

Table 5 illustrates this phenomenon even more starkly. The overall best method, Rankit, has its least accurate deviation value, 3.647, among small samples of the psychometric distribution, Mass at Zero with Gap. Van der Waerden attains its most accurate deviation value, 0.195, among large samples of the Digit Preference achievement distribution. The best method's worst deviation value on any distribution is 19 times higher than the worst method's best value. This pattern holds independently for sample size and distribution. Van der Waerden's best deviation values are superior to Rankit's worst among small and large samples. Sample size exerts a strong enough influence to reverse the standing of the best- and worst-performing methods on every distribution. All four methods perform best with Digit Preference and Multimodal Lumpy and worst with Mass at Zero with Gap.

Separately, the influence of sample size and distribution can make the worst normalization method outperform the best one. Together, their influence can distort the standard deviation enough to render the *T* score distribution, and the test results, meaningless. In the best case scenario, Rankit would be used among large samples of the Digit Preference distribution, where it is off target by 0.043 (Table 5). With a *Z* score of 2 and a standard error of ± 2, this leads to a true score range of 4.172, only 4% greater than the expected score range. In the worst case scenario, Van der Waerden could be used among small samples of the Mass at Zero with Gap distribution, where it is off target by 4.117. With the same *Z* score and standard error, this combination produces a true score range of 20.468, or 512% greater than the expected score range. Clearly, a true score range of 20 is psychometrically unacceptable.Telling a parent that her child scored somewhere between a 60 and an 80 is equally pointless.

Magnitude of Deviation on the Standard Deviation

Returning to the second accuracy measure, magnitude of deviation, Table 4 shows how consistently the methods perform on the standard deviation.[1] Among achievement distributions, they exhibit virtually no variability with large samples (RMS = 0.001) and slight variability with small samples (average RMS = 0.015). Among psychometric distributions, the pattern is the same but the magnitude of deviation is greater for both large and small samples (average RMS = 0.094 and 0.529, respectively). As expected, small samples and psychometric distributions aggravate the instability of each method's performance and exacerbate the differences between them. Average magnitude of deviation for small samples is nearly six times greater than larger samples. Average magnitude of deviation for psychometric distributions is 39 times greater than achievement distributions. Table 5 provides RMS values for all eight distributions. It is notable that Extreme Asymmetric – Growth, which is highly skewed, presents the highest RMS value among achievement distributions, although it is still lower than the psychometric distributions.

The Blom, Tukey, Van der Waerden, and Rankit approximations display considerable inaccuracy on the standard deviation, which has practical implications for test scoring and interpretation. Overestimation or underestimation of the standard deviation can bias comparisons of test-takers and tests. Therefore, practitioners should consider both sample size and distribution when selecting a normalizing procedure.

Small samples and skewed distributions aggravate the inaccuracy of all ranking methods, and these conditions are common in achievement and psychometric test data. However, substantial differences between methods are found among large samples and relatively symmetrical distributions as well. Therefore, scores from large samples should be plotted to observe population variance, in addition to propensity scores, tail weight, modality, and symmetry. Practitioners including analysts, educators, and administrators should also be advised that most test scores are less accurate than they appear. Caution should be exercised when making decisions based on standardized test performance.

This experiment demonstrates that Rankit is the most accurate method on the standard deviation when sample size and distribution are not taken into account; it is the most accurate method among both small and large samples; and it is the most accurate method among both achievement and psychometric distributions. Van der Waerden's approximation consistently performs the worst across sample sizes and distributions. In most cases, Blom's method comes in second place and Tukey's, third.

It would be useful to perform a more exhaustive empirical study of these ranking methods to better describe their patterns. It would also be of theoretical value to analyze the mathematical properties of their differences. More research can be done in both theoretical and applied domains. However, these results identify clear patterns that should guide the normalization of test scores in the social and behavioral sciences.

Note

[1]Curiously, the worst RMS values belong to Blom (Table 4), yet Blom achieves the second place deviation value on three out of four moments, among small and large samples and all eight distributions (Table 3). This suggests that Blom's approximation may achieve some technical precision at the expense of stability.

References

Aiken, L. R. (1987). Formulas for equating ratings on different scales. *Educational and Psychological Measurement*, *47*(1), 51-54.

Allport, F. M. (1934). The J-curve hypothesis of conforming behavior. *Journal of Social Psychology*, *5*, 141-183.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location survey and advances*. Princeton, NJ: Princeton University Press.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Bliss, C. I., Greenwood, M. L., & White, E. S. (1956). A Rankit analysis of paired comparisons for measuring the effect of sprays on flavor. *Biometrics*, *12*(4), 381-403. Retrieved March 26, 2007 from JSTOR database.

Blom, G. (1954). Transformation of the binomial, negative binomial, Poisson and χ^2 distributions. *Biometrika*, *41*(3/4), 302-316.

Blom, G. (1958). *Statistical estimates and transformed beta-variables*. NY: John Wiley & Sons.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Chang, S. W. (2006). Methods in scaling the basic competence test. *Educational and Psychological Measurement*, *66*, 907-929

Conover, W. J. (1980). *Practical nonparametric statistics*. NY: John Wiley & Sons.

Cronbach, L. J. (1976). *Essentials of psychological testing* (*3rd Ed.*). NY: Harper & Row.

Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale: Lawrence Erlbaum Associates.

Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver and Boyd.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675-701.

Gosset, W. S. ("Student") (1908). The probable error of a mean. *Biometrika*, *6*(1), 1-25.

Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika*, *48*(1/2), 151-165. Retrieved August 3, 2007 from JSTOR database.

Horst, P. (1931). Obtaining comparable scores from distributions of dissimilar shape. *Journal of the American Statistical Association*, *26*(176), 455-460. Retrieved August 23, 2007 from JSTOR database.

Ipsen, J., & Jerne, N. (1944). Graphical evaluation of the distribution of small experimental series. *Acta Pathologica, Microbiologica et Immunologica Scandinavica*, *21*, 343-361.

Kline, P. (2000). *Handbook of psychological testing* (*2nd Ed.)*. London: Routledge.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (*2nd Ed.*). NY: Springer Science+Business Media.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden-Day.

Lester, P. E., & Bishop, L. K. (2000). *Handbook of tests and measurement in education and the social sciences* (*2nd Ed.*). Lanham, MD: Scarecrow Press.

McCall, W. A. (1939). *Measurement*. NY: MacMillan.

Mehrens, W. A., & Lehmann, I. J. (1980). *Standardized tests in education* (*3rd Ed.*). NY: Holt, Rinehart and Winston.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.

Micceri, T. (1990). Proportions, pitfalls and pendulums. *Educational and Psychological Measurement*, *50*(4), 769-74.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, *3*(1), 55-67.

Nunnally, J. C. (1978). *Psychometric theory*. NY: McGraw-Hill.

Osborne, J. W. (2002). Normalizing data transformations. *ERIC Digest*, ED470204. Available online: www.eric.ed.gov

Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Societ*y, *Series A*, *186*, 343-414.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of a population distribution and the robustness of four simple test statistics. *Biometrika*, *62*, 223-241.

The Psychological Corporation. (1955). Methods of expressing test scores. *Test Service Bulletin*, *48*, 7-10.

Sawilowsky, S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education data sets. *Psychometrika*, *55*: 729.

Sawilowsky, S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*(2), 352-360.

Sawilowsky, S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with Fortran*. Oak Park: JMASM.

Solomon, S. R. (2008). *A comparison of ranking methods for normalizing scores*. Ph.D. dissertation, Wayne State University, United States - Michigan. Retrieved February 27, 2009, from Dissertations & Theses @ Wayne State University database. (Publication No. AAT 3303509).

SPSS (2006). *Statistical Package for the Social Sciences (SPSS)* 15.0 for Windows. Author.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*(6), 1055-1098.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departures from normality. *Communications in Statistics*, *A11*, 2485-2511.

Tapia, R. A., & Thompson, J. R. (1978). *Nonparametric probability density estimation*. Baltimore: Johns Hopkins University Press.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston, MA: Houghton Mifflin.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1-67. Retrieved August 3, 2007 from JSTOR database.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Indian Journal of Statistics*, *25*, 331-351.

Van der Waerden, B. L. (1952/1953a). Order tests for the two-sample problem and their power. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A), 55 (*Indagationes Mathematical 14*), 453-458, & 56 (*Indagationes Mathematicae 15*), 303-316.

Van der Waerden, B. L. (1953b). Testing a distribution function. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen (A),*56* (*Indagationes Mathematicae 15*), 201-207.

Walker, H. M., & Lev, J. (1969). *Elementary statistical methods* (*3rd Ed.*). NY: Holt, Rinehart and Winston.

Wilson, E. B., & Hilferty, M. M. (1929). Note on C. S. Peirce's experimental discussion of the law of errors. *Proceedings of the National Academy of Science*, *15*, 120-125.

Wimberley, R. C. (1975). A program for the T-score normal standardizing transformation. *Educational and Psychological Measurement*, *35*, 693-695.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 463-468

1538 – 9472/09/$95.00

Relationship between Internal Consistency and Goodness of Fit Maximum Likelihood Factor Analysis with Varimax Rotation

Gibbs Y. Kanyongo James B. Schreiber
Duquesne University

This study investigates how reliability (internal consistency) affects model-fitting in maximum likelihood exploratory factor analysis (EFA). This was accomplished through an examination of goodness of fit indices between the population and the sample matrices. Monte Carlo simulations were performed to create pseudo-populations with known parameters. Results indicated that the higher the internal consistency the worse the fit. It is postulated that the observations are similar to those from structural equation modeling where a good fit with low correlations can be observed and also the reverse with higher item correlations.

Key words: Factor structure; Matrices; Scree plot; Parallel Analysis.

Introduction

The purpose of the study is to investigate how reliability (internal consistency) affects model-fitting in maximum likelihood exploratory factor analysis (EFA). The study seeks to accomplish this through integrating and extending the work of Kanyongo (2006) on reliability and number of factors extracted and Fabringer, Wegener, MacCallum, Strahan's (1999) work on model-fitting and the number of factors extracted in exploratory factor analysis.

Internal Consistency

Henson (2001) noted that reliability is often a misunderstood measurement concept. There are several forms of reliability coefficients, but some of the most commonly used are internal consistency estimates. Internal consistency estimates relate to item homogeneity, or the degree to which the items on a test jointly measure the same construct (Henson, 2001). Thus, in the classical test theory, reliability is concerned with score consistency. The classical conceptualization of score reliability relates the concept of score consistency to true scores. Allen and Yen (1979) defined a person's true score as the theoretical average obtained from an infinite number of independent testings of the same person with the same test.

Many authors conceptualize three sources of measurement error within the classical framework: content sampling of items, stability across time, and interrater error (Henson, 2001). Content sampling refers to the theoretical idea that the test is made up of a random sampling of all possible items that could be on the test. If that is the case, the items should be highly interrelated because they assess the same construct of interest. This item interrelationship is typically called internal consistency, which suggests that the items on a measure should correlate highly with each other if they truly represent appropriate content sampling (Henson, 2001). If the items are highly correlated, it is theoretically assumed that the construct of interest has been measured to some degree of consistency, that is, the scores are reliable.

Internal consistency estimates are intended to apply to test items assumed to represent a single underlying construct, thus, the

Gibbs Y. Kanyongo is an Assistant Professor in the Department of Educational Foundations and Leadership in the School of Education. Email: kanyongog@duq.edu. James B. Schreiber is an Associate Professor in the Department of Foundations and Leadership in the School of Education. Email: schreiberj@duq.edu.

use of these estimates with speeded tests is inappropriate due to the confounding of construct measurement with testing speed. Furthermore, for tests that consist of scales measuring different constructs, internal consistency should be assessed separately for each scale (Henson, 2001).

Exploratory Factor Analysis (EFA)

The primary purpose of EFA is to arrive at a more parsimonious conceptual understanding of a set of measured variables by determining the number and nature of common factors needed to account for the pattern of correlations among the measured variables (Fabringer, et al., 1999). EFA is based on the common factor model (Thurstone, 1947). The model postulates that each measured variable in a battery of measured variables is a linear function of one or more common factors and one unique factor.

Fabringer, et al. (1999) defined common factors as unobservable latent variables that influence more than one measured variable in the battery and accounts for the correlations among the measured variables, and unique factors as latent variables that influence only one measured variable in a battery and do not account for correlations among measured variables. Unique factors are assumed to have two components; a specific component and an error of measurement component, the unreliability in the measured variable. The common factor model seeks to understand the structure of correlations among measured variables by estimating the pattern of relations between the common factor(s) and each of the measured variables (Fabringer, et al., 1999).

Previous Work

Kanyongo investigated the influence of internal consistency on the number of components extracted by various procedures in principal components analysis. Internal consistency reliability coefficients are not direct measures of reliability, but are theoretical estimates based on classical test theory. IC addresses reliability in terms of consistency of scores across a given set of items. In other words, it is a measure of the correlation between subsets of items within an instrument.

The study employed the use of Monte Carlo simulations to generate scores at different levels of reliability. The number of components extracted by each of the four procedures, scree plot, Kaiser Rule, Horn's parallel analysis procedure and modified Horn's parallel analysis procedure was determined at each reliability level. In his study, Kanyongo (2006) found mixed results on the influence of reliability on the number of components extracted. However, generally, when component loading was high, an improvement in reliability resulted in improvement of the accuracy of the procedures especially for variable-to-component ratio of 4:1.

The Kaiser procedure showed the greatest improvement in performance although it still had the worst performance at any given reliability level. When the variable-to-component ratio was 8:1, reliability did not impact the performance of the scree plot, Horn's parallel analysis (HPA) or modified Horn's parallel analysis (MHPA) since they were 100% accurate at all reliability levels. When component loading was low, it was not clear what the impact of reliability was on the performance of the procedures.

The work of Fabringer, et al. (1999) involved an examination of the use of exploratory factor analysis (EFA) in psychological research. They noted that a clear conceptual distinction exists between principal factor analysis (PCA) and EFA. When the goal of the analysis is to identify latent constructs underlying measured variables, it is more sensible to use EFA than PCA. Also, in situations in which a researcher has relatively little theoretical or empirical basis to make strong assumptions about how many common factors exist or what specific measured variables these common factors are likely to influence, EFA is probably a more sensible approach than confirmatory factor analysis (CFA**).**

Fabringer, et al. (1999) pointed that in EFA; sound selection of measured variables requires consideration of psychometric properties of measures. When EFA is conducted on measured variables with low communalities, substantial distortion in results occurs. One of the reasons why variables may have low communalities is low reliability. Variance due to

random error cannot be explained by common factors; and because of this, variables with low reliability will have low communality and should be avoided.

Fabringer, et al. (1999) also noted that although there are several procedures that are available for model-fitting in EFA, the maximum likelihood (ML) method of factor extraction is becoming increasingly popular. ML is a procedure used to fit the common factor model to the data in EFA. ML allows for the computation of a wide range of indices of the goodness of fit of the model. ML also permits statistical significance testing of factor loadings and correlations among and the computation of confidence intervals for these parameters (Cudeck & O'Dell, 1994). Fabringer, et al. (1999) pointed out that the ML method has a more formal statistical foundation than the principal factors methods and thus provides more capabilities for statistical inference, such as significance testing and determination of confidence intervals.

In their work, Frabinger, et al. (1999) further stated that, ideally, the preferred model should not just fit the data substantially better than simple models and as well as more complex models. The preferred model should fit the data reasonably well in an absolute sense. A statistic used for assessing the fit of a model in ML factor analysis solutions is called a goodness of fit index.

There are several fit indices used in ML factor analysis and one of them is the likelihood ratio statistic (Lawley, 1940). If sample size is sufficiently large and the distributional assumptions underlying ML estimation are adequately satisfied, the likelihood ratio statistic approximately follows a Chi-square distribution if the specified number of factors is correct in the population (Fabringer, et al., 1999). They noted that, if this is not the case, a researcher should exercise caution in interpreting the results because a preferred model that fits the data poorly might do so and because the data do not correspond to assumptions of the common factor model.Alternatively, it might suggest the existence of numerous minor common factors. Fabringer, et al. also suggested that "with respect to selecting one of the major methods of fitting the common factor model in EFA (i.e., principal factors, iterated principal factors, maximum likelihood), all three are reasonable approaches with certain advantages and disadvantages. Nonetheless, the wide range of fit indexes available for ML EFA provides some basis for preferring this method" (p.283). Since ML EFA has potential to provide misleading results when assumptions of multivariate normality are severely violated, the recommendation is that the distribution of the measured variables should be examined prior to using the procedure. If non-normality is severe (skew>2; kurtosis>7), measured variables should be transformed to normalize their distributions (Curran, West & Finch, 1996).

Fabringer, et al. (1999) noted that the root mean square error of approximation (RMSEA) fit index and the expected cross-validation index (ECVI) provide a promising approach for assessing fit of a model in determining the number of factors in EFA. They recommended that "In ML factor analysis; we encourage the use of descriptive fit indices such as RMSEA and ECVI along with more traditional approaches such as the scree plot and parallel analysis" (p.283). Based on this recommendation, this study uses these fit indices along with the scree plot and parallel analysis to assess the accuracy of determining the number of factor at a given level of reliability.

Research Question

The main research question that this study intends to answer is: As the internal consistency of a set of items increases, does the fit of the data to the exploratory factor analysis improve? To answer this question, a Monte Carlo simulation study was conducted which involved the manipulation of component reliability ($\rho_{xx'}$) loading (a_{ij}), variable-to-component ratio ($p:m$). The number of variables (p) was made constant at 24 to represent a moderately large data set.

Methodology

The underlying population correlation matrix was generated for each possible p, $p:m$ and a_{ij} combination, and the factors upon which this population correlation matrix was based were independent of each other. RANCORR program

by Hong (1999) was used to generate the population matrix as follows.

The component pattern matrix was specified with component loading of .80 and variable-to-component ratio of 8:1. After specifying the component pattern matrix and the program was executed, a population correlation matrix is produced. After the population correlation matrix was generated as described in the above section, the MNDG program (Brooks, 2002) was then used to generate samples from the population correlation matrix. Three data sets for reliability of .60, .80, and 1.00, each consisting of 24 variables and 300 cases were generated. Based on the variable-to-component ratio of 8:1, each dataset had 3 factors built in.

Analysis

An exploratory factor analysis, maximum likelihood, varimax rotation and a three factor specification, was used for each of the three data sets; coefficient alpha = .6, .8, and 1.0. Two goodness-of-fit indices were chosen for this analysis RMSEA and ECVI. RMSEA was chosen because it is based on the predicted versus observed covariances which is appropriate given that nested models are not being compared.

Hu and Bentler (1999) suggested RMSEA <= .06 as the cutoff for a good model fit. RMSEA is a commonly utilized measure of fit, partly because it does not require comparison with a null model. ECVI was chosen because it is based on information theory; the discrepancy between models implied and observed covariance matrices: the lower the ECVI, the better the fit. Finally, the Chi-square and degrees of freedom are provided for each analysis. The data were also submitted to a PCA using the scree plot and parallel analysis to assess the accuracy of determining the number of common factors underlying the data sets.

Results

The results in Table 1 show that the two measures of goodness-of-fit used in this study (RMSEA) and (ECVI) both display the same pattern; the smaller the alpha, the better the model fit. The best fit was obtained for alpha of 0.6, RMSEA (0.025) and ECVI (1.44). As alpha increased from 0.6 to 1.0, both indices increased; an indication that the model fit became poor. Based on Hu and Bentler's (1999) recommendation that the cutoff for a good fit be RMSEA <= 0.06, results here show that only alpha of 0.6 had a good fit. The goodness-of-fit indices therefore suggest that the three-factor model is acceptable at alpha value of 0.6.

Table 1: Goodness-of-Fit Indices

Alpha	Chi-Square (df)	RMSEA	ECVI
.6	247.153 (207)	.025	1.44
.8	436.535 (207)	.061	2.07
1.0	736.385 (207)	.092	3.07

Along with goodness-of-fit-indices, the dataset with the best fit was submitted to principal components analysis through the scree plot and parallel analysis. Results of the scree plot analysis are displayed in Figure 1 while parallel analysis results are shown in Table 2. The scree plot shows a sharp drop between the third and fourth eigenvalues; an indication that there were three distinct factors in the data sets. These results confirm the three-factor model as the best model for these data.

To interpret results of parallel analysis, real data eigenvalues must be larger than random data eigenvalues for them to be considered meaningful eigenvalues. Table 2 shows that the first three eigenvalues expected for random data (1.55, 1.46 and 1.39) fall below the observed eigenvalues for all the three values of alpha. However, the forth eigenvalue of the random data (1.34) is greater than the observed eigenvalues of all the three alpha values. Again, the results further confirm the three-factor model as the best model.

Conclusion

Results in this study were inconsistent with our original ideas of the pattern of goodness of fit and internal consistency. It was anticipated that high internal consistency would yield a better fit.

Figure 1: Results of the Scree Plot

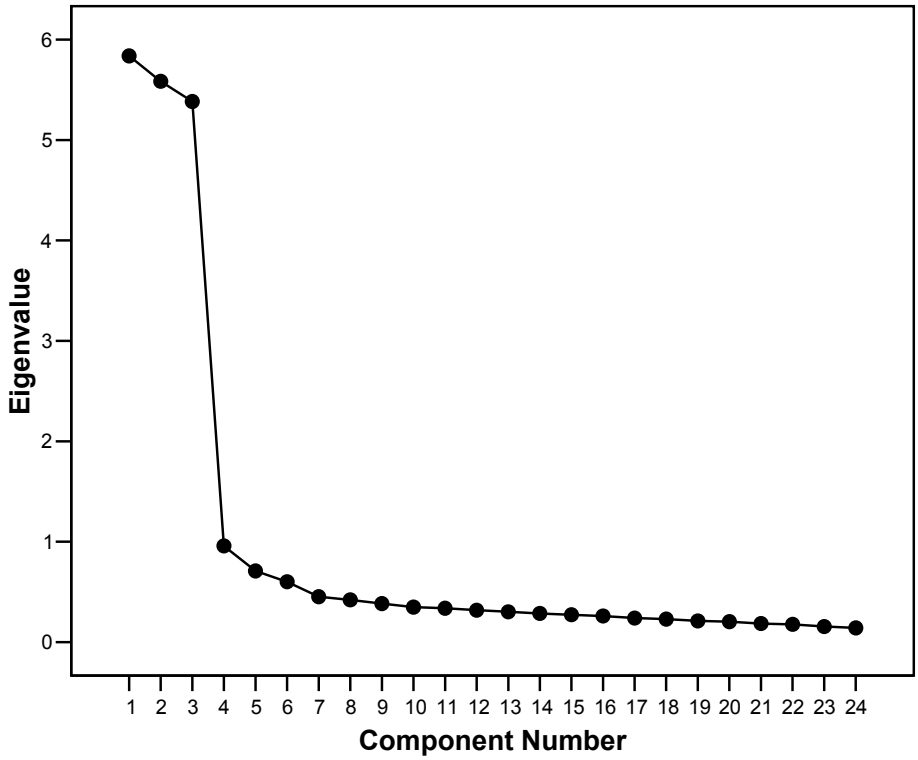


Table 2: Parallel Analysis Results

		Eigenvalues			
		1	2	3	4
Random Data		1.55	1.46	1.39	1.34
Real Data	.6	3.87	3.63	3.31	1.07
	.8	4.66	4.59	4.12	1.32
	1.0	5.84	5.58	5.38	.96

However, the findings seem logical because based on one author's experience with structural equation modeling, a perfect fit can exist when all variables in the model were completely uncorrelated if the variances are not constrained. Also, the lower the correlations stipulated in the model, the easier it is to find good fit. The stronger the correlations, the more power there is within structural equation modeling to detect an incorrect model. In essence, the higher the correlations, the more likely it is to incorrectly specify the model and observe a poor fit based on the indices. Second, if correlations are low, the researcher may lack the power to reject the model at hand.

Also, results seem to confirm what other researchers have argued in the literature. For example, Cudeck and Hulen (2001) noted that if a group of items has been identified as one-dimensional, the internal consistency of the

collection of items need not be high for factor analysis to be able to identify homogenous sets of items in a measuring scale. Test reliability is a function of items. Therefore, if only a few items have been identified as homogeneous by factor analysis, their reliability may not be high.

If ML with exploratory factor analysis including goodness-of-fit analyses are to be used more extensively in the future, a great deal of work must to be done to help researchers make good decisions. This assertion is supported by Fabringer, et al. (1999) who noted that, "although these guidelines for RMSEA are generally accepted, it is of course possible that subsequent research might suggest modifications" (p.280).

Limitations of Current Research

Since the study involved simulations, the major limitation of the study, like any other simulation study is that the results might not be generalizable to other situations. This is especially true considering the fact that the manipulation of the parameters for this study yielded strong internal validity thereby compromising external validity. However, despite this limitation, the importance of the findings cannot be neglected because they help inform researchers on the need to move away from relying entirely on internal consistency as a measure of dimensionality of data to an approach where other analyses are considered as well. This point was reiterated by Cudeck and Hulin (2001) who stated that a reliable test need not conform to a one-factor model and conversely items that fit a single common factor may have low reliability.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Brooks, G. P. (2002). *MNDG*. [http://oak.cats.ohiou.edu/~brooksg/mndg.htm]. (Computer Software and Manual).

Cudek, R., & Hulen, C. (2001). Measurement. *Journal of Consumer Psychology*, *10*, 55–69.

Cudek, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin*, *115*, 475-487.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16-29.

Fabringer, R. L., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments & Computers*, *31*, 727-730.

Hu, L., & Bentler, M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Kanyongo, G. Y. (2006). The influence of reliability of four rules for determining the number of components to retain. *Journal of Modern Applied Statistical Methods*, *2*, 332-343.

Lawrey, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinborough*, *60A*, 64-72.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 469-477

1538 – 9472/09/$95.00

Detecting Lag-One Autocorrelation in Interrupted Time Series Experiments with Small Datasets

Clare Riviello S. Natasha Beretvas
University of Texas at Austin

The power and type I error rates of eight indices for lag-one autocorrelation detection were assessed for interrupted time series experiments (ITSEs) with small numbers of data points. Performance of Huitema and McKean's (2000) z_{HM} statistic was modified and compared with the z_{HM}, five information criteria and the Durbin-Watson statistic.

Key words: Autocorrelation, information criteria, type I error, power.

Introduction

Educational research contains many examples of single-subject designs (Huitema, McKean, & McKnight, 1999). Single-subject designs, also known as interrupted time series experiments (ITSEs), are typically used to assess a treatment's effect on special populations such as children with autism or developmental disabilities (Tawney & Gast, 1984). The design consists of repeated measures on an outcome for an individual during baseline and treatment conditions (A and B phases, respectively). Use of repeated measures on an individual is designed such that the subject acts as his/her own control; this also helps rule out the possible influence of potential threats to validity including history, practice, and maturation effects.

With ITSE data, the pattern of scores over time is compared for the A (baseline) versus the B (treatment) phases. The comparison can lead to inferences about the effect of introducing the treatment on the trend in the outcome scores. To describe the change in trend, the effect on the level of the scores and on the possible growth pattern must be assessed. Numerical descriptors of these trends are not well estimated given the number of repeated measures is as small as is commonly found in educational single-case design research (Busk & Marascuilo, 1988; Huitema, 1985). One of the sources of these estimation problems is related to the autocorrelated structure inherent in such designs (Huitema & McKean, 1991; White, 1961; Kendall, 1954; Marriott & Pope, 1954).

Several test statistics and indices recommended for identifying potential autocorrelation exist. Unfortunately these statistics are typically recommended only for datasets with a larger numbers of data points than are typically encountered with ITSEs. Huitema and McKean (2000) introduced a test statistic, z_{HM}, to identify lag-one autocorrelation in small datasets. The Type I error rate of the z_{HM} was within nominal levels and sufficient power was associated with this statistic. The current study introduces a modification of the z_{HM} designed to enhance further its statistical power. This study assesses the Type I error rate and power of both versions of the z_{HM}. The performance of the two z_{HM} statistics is also compared with that of other test statistics and indices that are commonly used to identify autocorrelated residuals for models used to summarize trends for small ITSE datasets.

Clare Riviello is an Engineering Scientist at Applied Research Laboratories. Email: clareriviello@gmail.com. Natasha Beretvas is an Associate Professor and chair of Quantitative Methods in the Department of Educational Psychology. Her interests are in multilevel and meta-analytic modeling techniques. Email: tasha.beretvas@mail.utexas.edu.

Autocorrelation

One of the fundamental assumptions when using ordinary least squares estimation for multiple regression is that errors are independent. When the independence assumption does not hold, this leads to inaccurate tests of the partial regression coefficients (Huitema and McKean, 2000). For data consisting of repeated measures on an individual, it is likely that a model can explain some but not all of the autocorrelation. In addition, when the residuals in a regression model are autocorrelated the model must account for this to ensure accurate and precise estimation of parameters and standard errors. Thus, it is important to be able to detect autocorrelation so that the proper methods for estimating the regression model can be employed.

This study is designed to focus solely on first-order (lag-one) autocorrelation. For a multiple regression model including k predictors, x_i, of outcome y at time t using:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ... \beta_k x_{kt} + \varepsilon_t . \quad (1)$$

If there is a lag-one autocorrelation, ρ_1, between residuals, then ε_t, the residual at time t, is related to ε_{t-1}, the residual at time $t-1$ as follows:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + a_t \quad (2)$$

where ρ_1 is the autocorrelation between residuals separated by one time period. It is assumed that ε_t and ε_{t-1} have the same variance, and a_t is assumed to follow a standard normal distribution.

Estimating Lag-One Autocorrelation

Several formulas are available for estimating the lag-one correlation coefficient, ρ_1, for a time series consisting of N data points. The conventional estimator is calculated as follows:

$$r_1 = \frac{\sum_{t=1}^{N-1} (Y_t - \bar{Y})(Y_{t+1} - \bar{Y})}{\sum_{t=1}^{N} (Y_t - \bar{Y})^2} \quad (3)$$

where $\bar{Y}$ is the simple average of the N values of y. Unfortunately, as evidenced by its common usage, the bias of r_1 is often ignored. The expected value of a lag-1 autocorrelation coefficient for a series consisting of N data points was analytically derived by Marriott and Pope (1954) to be:

$$E(r_1) = \rho_1 - \frac{1}{N}(1 + 3\rho_1) + O(N^{-2}) . \quad (4)$$

It should be noted that the expression in Equation 2 only covers terms to order N^{-1} [thus, the term: $O(N^{-2})$]; there are additional terms for higher orders of the inverse of N. For large samples, these higher order terms tend towards zero. However, the ITSEs of interest in this study tend to involve short series where N is reasonably small and these higher order terms are thus not as negligible. Bias clearly exists in the estimation of the autocorrelation.

Huitema and McKean (1991) listed four additional, fairly common estimators designed to reduce the bias observed in r_1. However, each of these is also highly biased for small data sets. Huitema and McKean (1991) suggested correcting for the bias in r_1 by using

$$r_1^+ = r_1 + \frac{1}{N} \quad (5)$$

which, for smaller true values of ρ_1 incorporates some of the noted bias evident in Equation 2. The authors showed that their modified estimator, r_1^+, is unbiased when ρ_1 equaled zero even for sample sizes as small as $N = 6$. Additionally, the authors found that the bias was lower for positive values of ρ_1 but higher for some negative values.

When estimating the autocorrelation, it is also necessary to calculate the error variance

of the estimator because the estimator and its variance can be combined to produce a statistic that can be used to statistically test for the autocorrelation. Bartlett (1946) derived his variance formula for the variance of r_1:

$$\sigma_{r_1}^2 = \frac{1-\rho_1^2}{N}. \quad (6)$$

by ignoring terms of order N^{-2} or higher. This formula is commonly reduced to:

$$\hat{\sigma}_{r_1}^2 = \frac{1}{N} \quad (7)$$

under the assumption of the null hypothesis that $\rho_1 = 0$ (Huitema & McKean, 1991). Huitema and McKean (1991) asserted that the commonly used Bartlett variance approximation is not satisfactory for small sample sizes. Their simulation study indicated that $\hat{\sigma}_{r_1}^2$ (see Equation 7) consistently overestimated the empirical variance. This overestimation performed quite badly for values of N of less than twenty with Bartlett's variance approximation exceeding the empirical variance by 83% and 40% for $N = 6$ and $N = 10$, respectively. The authors explored the performance of Moran's variance estimate:

$$\hat{\sigma}_{r_1}^{2*} = \frac{(N-2)^2}{N^2(N-1)} \quad (8)$$

which, under the null hypothesis ($\rho_1 = 0$), gives precise error variance estimates. After looking at the performance of an autocorrelation test statistic using $\hat{\sigma}_{r_1}^{2*}$ as the error variance estimator, the authors concluded that $\hat{\sigma}_{r_1}^{2*}$ was not adequate for small sample sizes. In tests for positive values of autocorrelation, its results were too conservative except for large values of N. They recommended using:

$$\hat{\sigma}_{r_1}^{2+} = \frac{(N-2)^2}{N^2(N-1)}\{1-[E(r_1)]^2\} \quad (9)$$

where

$$E(r_1) \cong \rho_1 - \frac{1}{N}(1+3\rho_1). \quad (10)$$

(Marriot & Pope, 1954) as follows from Equation 4. Use of Equation 9 yielded values close to the empirical values of the variance of ρ_1 estimates even for Ns as small as $N = 6$.

Detecting Autocorrelation

The main purpose of estimating the correlation coefficient and calculating its error variance is to detect the presence of autocorrelation in a data set. If data are known to be autocorrelated, then methods other than ordinary least squares should be used to more accurately estimate the regression coefficients and their standard errors. One of the more commonly used tests for autocorrelation in residuals is the Durbin-Watson test statistic:

$$d = \frac{\sum_{t=2}^{N}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{N}\varepsilon_t^2} \quad (11)$$

where ε_t represents the residual at time t (see Equation 2).

The procedure for carrying out this test can be confusing, thus the sequence of steps for testing the non-directional $H_0 : \rho_1 = 0$ is explained here. First both d and (4–d) should be compared with the upper bound d_u. If both exceed this bound, then the null hypothesis is retained; otherwise, both d and (4–d) are compared with the lower bound, d_l. If either falls below d_l, then the null hypothesis is rejected and a non-zero lag one autocorrelation is inferred. If neither d nor (4–d) falls below d_l, the test is inconclusive. The concept of an inconclusive region is unsettling and, although computer methods that provide exact p-values are now becoming available, most are slow or expensive (Huitema & McKean, 2000).

It is in this context, that Huitema and McKean (2000) proposed an alternative test statistic that is simple to compute,

approximately normally distributed and does not have an inconclusive region. The test statistic was evaluated for its use to test residuals from ITSE models that have one to four phases. Huitema and McKean's test statistic is defined as:

$$z_{HM} = \frac{r_1 + \frac{P}{N}}{\sqrt{\frac{(N-2)^2}{N^2(N-1)}}} \quad (12)$$

where P is the number of parameters in the time-series regression model and N is the total number of observations in the time series. The authors found that

$$r_{1,P}^{+} = r_1 + \frac{P}{N} \quad (13)$$

provided an unbiased estimate of ρ_1 and that the denominator of the test statistic (in Equation 12) approximates the empirical variance of $r_{1,P}^{+}$ (see Equation 8).

The z_{HM} test statistic is a generalization of the test proposed in Huitema and McKean's (1991) earlier work was designed for a single-phase model of ITSE data. However, the authors failed to implement all of the suggestions from their previous study. Specifically, the authors did not use the corrected error variance, $\hat{\sigma}_{r_1}^{2+}$, (see Equation 9) that they had recommended. Instead they used $\hat{\sigma}_{r_1}^{2*}$ (see Equation 8). Because $\{1-[E(r_1)]^2\} \leq 1$, use of $\hat{\sigma}_{r_1}^{2+}$ should lead to a smaller variance and thus a larger value of the test statistic and increased power over $\hat{\sigma}_{r_1}^{2*}$.

Information Criteria

As an alternative to using test statistics to detect autocorrelated residuals, it is also possible to estimate a model using ordinary least squares regression, estimate the same model assuming autocorrelated residuals, and then compare the fit of the two models. A post-hoc evaluation that compares the two models' fit can be then be conducted using an information criterion such as Akaike's Information Criterion (AIC):

$$AIC = -2Log(L) + 2k \quad (14)$$

where L is the value of the likelihood function evaluated for the parameter estimates and k is the number of estimated parameters in a given model. The model with the smallest information criterion value is considered the best fitting model.

As an alternative to the asymptotically efficient but inconsistent AIC, several more consistent model fit statistics have been proposed (Bozdogan, 1987; Hannon & Quinn, 1979; Hurvich &Tsai, 1989; Schwarz, 1978). These include Swartz's (1978) Bayesian criterion:

$$SBC = -2Log(L) + Log(N)k \quad (15)$$

where N is the number of observations, Hannon and Quinn's (1979) information criterion

$$HQIC = -2Log(L) + 2kLog(Log(N)); \quad (16)$$

and Bozdogan's (1987) consistent AIC

$$CAIC = -2Log(L) + (Log(N)+1)k. \quad (17)$$

In addition, Hurvich and Tsai (1989) developed a corrected AIC specifically for small sample sizes, which deals with AIC's tendency to overfit models:

$$AICC = -2Log(L) + \frac{2kN}{N-k-1}. \quad (18)$$

For each of these information criteria formulations, the smaller the value, the better the model fit.

The AIC and SBC are supplied by default by most statistical software. For example, when using SAS's PROC AUTOREG (SAS Institute Inc., 2003) to estimate an autoregressive model, the procedure also provides results under the assumption of no

autocorrelation in residuals (i.e., using ordinary least squares, OLS, estimation). The procedure automatically provides the AIC and SBC for the OLS and autoregressive models to enable a comparison of the fit of the two models. To date, no studies have been conducted to compare use of information criteria for identification of autocorrelated residuals for ITSE data with small sample sizes.

Research Question

This study is designed to introduce and evaluate use of the variance correction suggested by Huitema and McKean (1991) in a modified version of their test statistic, z_{HM}. Specifically, the corrected test statistic being suggested and evaluated is:

$$z^{+}_{HM} = \frac{r_1 + \frac{P}{N}}{\{1-[E(r_1)]^2\}\sqrt{\frac{(N-2)^2}{N^2(N-1)}}} \qquad (19)$$

Identification of lag-one autocorrelation (of residuals) was compared for the z^{+}_{HM} and z_{HM} test statistics, the Durbin-Watson test statistic and the AIC, SBC, HQIC, CAIC, and AICC fit indices for conditions when $\rho_1 = 0$ and when $\rho_1 \neq 0$. This study focused only on two-phase ITSE data. This design lies at the root of commonly used single-subject designs and provides an important starting point for this investigation.

Methodology

SAS code was used to generate data, estimate models, and summarize results (Fan, Felsovalyi, Keenan, & Sivo, 2001). Several design conditions were manipulated to assess their effect on the performance of the test statistics and fit indices. These conditions included the magnitude of the treatment's effect on the level and linear growth, the degree of autocorrelation and the overall sample size of the ITSE data.

Model and Assumptions

The following two-phase, ITSE model (Huitema & McKean, 1991) was used to generate the data:

$$y_t = \beta_0 + \beta_1 t_t + \beta_2 d_t + \beta_3 [t_t - (n_A + 1)]d_t + \varepsilon_t \qquad (20)$$

where n_A is the number of data points in the first phase (baseline phase A), d_t is the dummy variable coded with a zero for data points in the baseline phase and with a one for data points in the second phase, and $[t_t - (n_A + 1)]d_t$ is the centered interaction between time and treatment. The interaction term is centered in this way to provide a coefficient, β_3, that represents the treatment's effect on the slope (i.e., the difference in the linear growth between that predicted using the treatment phase data and that predicted using the baseline data). The coefficient, β_2, represents the change in the intercept from the baseline to the treatment phase (specifically, the difference in the value of y_t, when $t = n_A + 1$, predicted using treatment versus baseline phase data).

Thus, the β_2 and β_3 coefficients describe the effect of the treatment on the level and growth in y, respectively. The residuals (ε_t) were generated such that $\varepsilon_t = \rho_1 \varepsilon_{t-1} + a_t$ with ρ_1 being the true lag-one autocorrelation between residuals separated by one time unit, and a_t was randomly and independently selected from a standard normal distribution.

Because the focus in ITSE designs is on the effect of the intervention, the β_2 and β_3 coefficients (see Equation 20) are of most interest. Thus, when generating the data in this simulation study, values of β_0 (baseline data's intercept) and of β_1 (baseline data's linear growth) were not manipulated but were fixed such that β_0 was set to zero and β_1 was set to a value of 0.2 in all scenarios. This modeled data with an intercept of zero (i.e., $y_t = 0$ at $t = 0$) and a slight baseline trend. Values of β_2 and β_3, however, were varied to investigate their effect on detecting autocorrelation. Each parameter took on values 0, 0.2, and 0.4 in this fully crossed design.

In order to evaluate how the model selection criteria performed over the range of possible values for ρ_1, its value was varied to range from −0.8 up to 0.8 in increments of 0.2.

Finally, the number of data points, *N*, in the two phases for each scenario were varied to be 12, 20, 30, 50, or 100 with the points being divided equally between the two phases so that $n_A = n_B$ with values for each of: 6, 10, 15, 25, or 50.

The simulation study thus entailed a fully crossed design consisting of three values of β_2 crossed with three values of β_3, crossed with nine values of ρ_1, crossed with five values of *N* for a total of 405 combinations of conditions. One thousand datasets were generated for each of these 405 scenarios.

Analyses

After each dataset was generated, the regression model in Equation 20 was estimated using SAS's PROC AUTOREG. This procedure estimates the model using both ordinary least squares (OLS) (assuming $\rho_1 = 0$) and autoregressive methods (assuming $\rho_1 \neq 0$). The procedure provides values for the AIC and SBC for both models. HQIC, CAIC, and AICC were then calculated (see Equations 16, 17 and 18, respectively) using the log likelihood obtained from the AIC value. For each information criterion, a tally was kept describing when the autoregressive model's information criterion was lower than that of the OLS model. PROC AUTOREG additionally provides the *p*-value for the Durbin-Watson test statistic. As with the AIC and SBC, a tally was kept of the proportion of trials for which this *p*-value led to a rejection of the null hypothesis that $\rho_1 = 0$ ($p < .05$).

The z_{HM} and z^+_{HM} were also calculated (see Equation 12 and 19, respectively) using the residuals from the OLS regression. The $E(r_1)$ in the denominator of Equation 19 was obtained by substituting $r^+_{1,P}$ for the unknown ρ_1 in Equation 6. Again, a tally was kept describing the proportion of trials for which the null hypothesis of no autocorrelation was rejected ($p < .05$). For conditions in which $\rho_1 \neq 0$, the tally by scenario for each of the eight model selection criteria provided the power to identify the correct model. For conditions in which $\rho_1 = 0$, the tally provided the type I error rate.

Results

Type I Error Rates

Table 1 contains Type I error rates by condition and criterion. Sample size appeared to have the strongest effect on type I error rates. The type I error rate was not greatly affected by the values of β_2 and β_3. Overall, the Type I error rates for z_{HM} and z^+_{HM} were the best of the eight criteria investigated. The rates were somewhat conservative for the smallest sample size conditions ($N = 12$) with values of 0.022 and 0.035 for z_{HM} and z^+_{HM}, respectively. The z_{HM} maintained type I error rates at the nominal level across sample size conditions (with a maximum value of 0.051). The rates for z^+_{HM} were slightly elevated (with values of 0.059) although the statistic performed much better than did the Durbin-Watson (DW) and the five information criteria (ICs) investigated.

The Type I error rates of the five ICs (SBC, AIC, HQIC, CAIC and AICC) and for the DW statistic were generally inflated across the $\rho_1 = 0$ conditions examined with the indices performing from worst to best as follows: AIC, HQIC, SBC, AICC, DW, CAIC. The Type I error rate inflation, however, decreased with increasing sample size. Only in the scenarios with the largest sample size ($N = 100$), were the CAIC and SBC's Type I error rates acceptable if somewhat conservative. The CAIC's Type I error rate performance was also acceptable (0.056) for conditions in which *N* was 50.

Power

Table 2 displays the power of the eight criteria used to evaluate the presence of lag-one autocorrelated residuals. In the presence of type I error inflation, the power of a criterion becomes somewhat moot. Thus, it should be kept in mind that the Type I error inflation noted for the DW and the five ICs. As would be expected, for all criteria the power was found to increase for larger sample sizes. Similarly, it was expected and found that as the magnitude of ρ_1 increased so did the power to detect the ρ_1 of the ICs and test statistics. The z_{HM} and z^+_{HM} exhibited consistently better power levels than the SBC and DW for all positive values of ρ_1.

Table 1: Type I Error Rates (False Detection) of Lag-One Autocorrelation by Criterion and Condition

Condition		Information Criterion					Test Statistics ($p < .05$)		
Parm*	True Value	SBC	AIC	HQIC	CAIC	AICC	DW	z_{HM}	z^{+}_{HM}
ρ_1	0	0.185	0.304	0.264	0.129	0.168	0.146	0.043	0.053
β_2	0.4	0.185	0.303	0.265	0.128	0.172	0.143	0.044	0.054
	0.2	0.185	0.303	0.262	0.129	0.167	0.145	0.043	0.053
	0	0.185	0.305	0.264	0.129	0.166	0.149	0.042	0.052
β_3	0.4	0.188	0.305	0.266	0.131	0.172	0.147	0.044	0.055
	0.2	0.187	0.306	0.264	0.129	0.165	0.147	0.043	0.052
	0	0.180	0.300	0.262	0.127	0.168	0.143	0.042	0.051
N	12	0.424	0.490	0.523	0.316	0.131	0.173	0.022	0.035
	20	0.228	0.343	0.316	0.155	0.182	0.164	0.047	0.059
	30	0.146	0.272	0.221	0.092	0.183	0.149	0.047	0.059
	50	0.087	0.225	0.157	0.056	0.178	0.132	0.051	0.058
	100	0.038	0.190	0.103	0.024	0.167	0.110	0.049	0.052

*Parm. = Parameter

Both of these test statistics had better power than all other indices when $\rho_1 \geq 0.6$. These results also supported the theoretical conclusion mentioned earlier that z^{+}_{HM} will always have more power than z_{HM}. For negative values of ρ_1, the ICs and DW statistic exhibited better power than the z_{HM} and z^{+}_{HM}. And the ICs that performed worst in terms of type I error control performed best in terms of power.

The power was also unaffected by the true values of β_2 and β_3. The power of z_{HM} and z^{+}_{HM} was quite low (0.089 and 0.133, respectively) for the $N = 12$ conditions but the power levels become more comparable to those of the other criteria for larger *N*. However, only z_{HM} and z^{+}_{HM} had exhibited acceptable type I error rates.

Conclusion

The results of the simulation study support use of the z_{HM} and z^{+}_{HM} for identification of lag-one autocorrelation in small ITSE datasets. Both statistics maintain nominal rates of type I error control although z^{+}_{HM}'s rates seemed slightly inflated in the larger sample size conditions. Concomitant with the type I error control were found somewhat lower empirical power levels. However the type I error inflation of the five ICs and the DW prohibit their use for detection of autocorrelation in the conditions examined here and especially with ITSE data consisting of a small number of data points.

A type I error in the current context means that an autoregressive model will be estimated unnecessarily. While this should have minimal effect on the estimation of the β coefficients in Equation 20, it will likely affect the standard error (*SE*) estimates used to test the

Table 2: Power to Detect Lag-One Autocorrelation by Criterion and Condition

Condition		Information Criterion					Test Statistic ($p < .05$)		
Parm*	True Value	SBC	AIC	HQIC	CAIC	AICC	DW	z_{HM}	z^{+}_{HM}
ρ_1	0.8	0.614	0.672	0.661	0.569	0.585	0.574	0.689	0.699
	0.6	0.530	0.609	0.594	0.480	0.516	0.500	0.621	0.633
	0.4	0.380	0.494	0.462	0.320	0.392	0.370	0.476	0.492
	0.2	0.169	0.299	0.258	0.120	0.194	0.164	0.204	0.218
	-0.2	0.499	0.670	0.616	0.399	0.503	0.473	0.188	0.212
	-0.4	0.830	0.894	0.883	0.765	0.769	0.765	0.489	0.526
	-0.6	0.952	0.970	0.968	0.926	0.896	0.904	0.697	0.734
	-0.8	0.988	0.992	0.993	0.981	0.963	0.968	0.830	0.865
β_2	0.4	0.622	0.702	0.679	0.571	0.603	0.591	0.526	0.549
	0.2	0.619	0.699	0.681	0.569	0.601	0.590	0.523	0.546
	0	0.619	0.699	0.678	0.570	0.602	0.588	0.523	0.546
β_3	0.4	0.622	0.701	0.680	0.570	0.602	0.590	0.524	0.547
	0.2	0.620	0.700	0.680	0.570	0.603	0.590	0.524	0.548
	0	0.618	0.699	0.679	0.570	0.602	0.589	0.525	0.547
N	12	0.515	0.560	0.579	0.440	0.287	0.323	0.089	0.133
	20	0.461	0.544	0.524	0.404	0.424	0.415	0.377	0.412
	30	0.571	0.670	0.636	0.515	0.605	0.570	0.564	0.585
	50	0.717	0.812	0.775	0.678	0.788	0.754	0.732	0.743
	100	0.836	0.914	0.883	0.813	0.908	0.887	0.860	0.863

*Parm. = Parameter

statistical significance of these coefficients. The current evaluation could be extended further by comparing estimation of the OLS versus autoregressive model coefficients and their *SE*s for different levels of autocorrelation. This could help inform the current study's type I error and power results by indicating the magnitude of the effect of incorrect modeling of autocorrelation. For example, if only a small degree of accuracy and precision is gained by modeling the autocorrelation for a certain value of ρ_1, then it may not matter that the model selection criteria has low power at that value. Similarly, if an insubstantial degree of accuracy and precision results from false identification of autocorrelation, then the type I error inflation noted in this study might be of minimal importance.

As with most simulation studies, results are limited by the conditions investigated: the values of the β_2 and β_3 coefficients (see Equation 20) do not seem to have much effect on identification of ρ_1, but it should be investigated whether this is really the case or whether it just appears that way from the limited range of values of β_2 and β_3 that were chosen in this study. One of the main limitations of this study is that it considers only the two-phase ITSE data and only investigated first-order autocorrelation. Another important limitation is that performance was evaluated only for a small subset of possible data trends. All conditions included a slight positive linear trend in

baseline. In addition, the only model misspecification assessed was whether the residuals were autocorrelated.

Future research should investigate use of the z_{HM} and z^{+}_{HM} for further misspecified models including when a true non-linear trend is ignored to mimic asymptotic trends resulting from ceiling or floor effects. The performance of these statistics could also be assessed for ITSEs with more than two phases (e.g., for ABAB designs) as investigated by Huitema and McKean (2000). This study also only investigated conditions in which the treatment and baseline phases had equal numbers of data points ($n_B = n_A$). Single-subject studies frequently entail unequal sample sizes per phase and the effect of uneven *n* should be investigated.

Based on the results of this study, researchers interested in modeling linear growth in ITSE data with a small number of data points should use z^{+}_{HM} or z_{HM} to test for the presence of lag-one autocorrelation. Researchers are cautioned against using the Durbin-Watson test statistic and the various information criteria evaluated here including the AIC, HQIC, SBC, AICC, DW and the CAIC for two-phase ITSEs with *N*s less than 50.

References

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society*, *8*, 27-41.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*, 229-242.

Fan, X., Felsovalyi, A., Keenan, S. C., & Sivo, S. (2001). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute, Inc.

Hannon, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, *Series B*, *41*, 190-195.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*, 107-118

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, *110*, 291-304.

Huitema, B. E., & McKean, J. W. (2000). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports*, *87*, 3-20.

Huitema, B. E., Mckean, J. W., & McKnight S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement*, *59*, 767-786.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297-307.

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika*, *41*, 403-404.

Marriott, F. H. C., & Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika*, *41*, 390-402.

SAS Institute, Inc. (2003). SAS (Version 9.1) [Computer Software]. Cary, NC: SAS Institute, Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Tawney, L., & Gast, D. (1984). *Single-subject research in special education*. Columbus, OR: Merrill.

White, J. S. (1961). Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika*, *48*, 85-94.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 478-487

1538 – 9472/09/$95.00

Estimating the Parameters of Rayleigh Cumulative Exposure Model in Simple Step-Stress Testing

Mohammed Al-Haj Ebrahem Abedel-Qader Al-Masri
Yarmouk University
Irbid, Jordan

Assumes the life distribution of a test unit for any stress follows a Rayleigh distribution with scale parameter θ, and that $Ln(\theta)$ is a linear function of the stress level. Maximum likelihood estimators of the parameters under a cumulative exposure model are obtained. The approximate variance estimates obtained from the asymptotic normal distribution of the maximum likelihood estimators are used to construct confidence intervals for the model parameters. A simulation study was conducted to study the performance of the estimators. Simulation results showed that in terms of bias, mean squared error, attainment of the nominal confidence level, symmetry of lower and upper error rates and the expected interval width, the estimators are very accurate and have a high level of precision.

Key words: Accelerated life test, Cumulative exposure model, Rayleigh distribution, Maximum likelihood estimation, Step-stress.

Introduction

The Rayleigh distribution arises in a variety of fields. This distribution is frequently employed by engineers and scientists as a model for data resulting from investigations involving wave propagation, radiation and related inquiries as well as in the analysis of target error data Cohen and Whitten (1988). Some types of electro vacuum devices age rapidly with time even though they may have no manufacturing defects, the Rayleigh distribution is quite appropriate for modeling the lifetime of such units as it possesses a linearly increasing hazard rate Johnson, Kotz and Balakrishnan (1994). Other applications and motivations for the Rayleigh distribution can be found in Cohen and Whitten (1988).

Mohammed Al-Haj Ebrahem is an Associate Professor in the Department of Statistics. Email: m_hassanb@hotmail.com. Abedel-Qader Al-Masri is an Instructor in the Department of Statistics. Email: almasri68@yahoo.com

Accelerated life tests are used to quickly obtain information on the life distribution of products by testing them at higher than nominal levels of stress to induce early failures. Data are obtained at accelerated conditions and based on a regression type model, results are extrapolated to the design stress to estimate the life distribution; such overstress testing reduces time and cost. One method of applying stress to the test units is a step-stress scheme which allows the stress of a unit to be changed at specified times. Nelson (1980) described this important type of accelerated life test. In step-stress testing, a unit is placed on a test at an initial low stress, if it does not fail in a predetermined time, τ, stress is increased. If there is a single change of stress, the accelerated life test is called a simple step-stress test.

The cumulative exposure model defined by Nelson (1990) for simple step-stress testing with low stress X_1 and high stress X_2 is:

$$G(t) = \begin{cases} G_1(t) & t \le \tau \\ G_2(t-\tau+s) & t > \tau \end{cases}$$

where $G_i(t)$ is the cumulative distribution function of the failure time at stress X_i, τ is the

time to change stress and s is the solution of $G_1(\tau) = G_2(s)$.

Most of the available literature on a step-stress accelerated life testing deals with the exponential exposure model. Khamis and Higgins (1996, 1998) proposed a new model known as KH model for step-stress accelerated life test as an alternative to the Weibull cumulative exposure model.

Miller and Nelson (1983) obtained the optimum simple step-stress accelerated life test plans for the case where the test units have exponentially distributed lifetimes. Bai, Kim and Lee (1989) extended the results of Miller and Nelson (1983) to the case of censoring. Khamis and Higgins (1996) obtained the optimum 3-step step-stress using the exponential distribution. Alhadeed and Yang (2005) obtained the optimum design for the lognormal step-stress model. Al-Haj Ebrahem and Al Masri (2007(a)) obtained the optimum simple step-stress plans for the log-logistic cumulative exposure model, by minimizing the asymptotic variance of the maximum likelihood estimate of a given 100 P-th percentile of the distribution at the design stress.

Al-Haj Ebrahem and Al Masri (2007(b)) obtained the optimum simple step-stress plans for the log-logistic distribution under time censoring. Xiong (1998) presented the inferences of parameters in the simple step-stress model in accelerated life testing with type two censoring. Xiong and Milliken (2002) studied statistical models in step-stress accelerated life testing when stress change time are random and obtained the marginal life distribution for test units. Nonparametric approaches for step-stress testing have been proposed by Shaked and Singurwalla (1983) and Schmoyer (1991). For additional details, see Chung and Bai (1998) and Gouno (2001). This article considers point and interval estimation of Rayleigh cumulative exposure model parameters.

Model and Assumptions

The probability density function and the cumulative distribution function of the Rayleigh distribution are given respectively by:

$$f(y,\theta) = \begin{cases} \frac{2y}{\theta} e^{\frac{-y^2}{\theta}} & , y > 0, \theta > 0 \\ 0 & , otherwise \end{cases} \quad (1)$$

and

$$F(y,\theta) = \begin{cases} 0 & , y < 0 \\ 1 - e^{\frac{-y^2}{\theta}} & , y > 0 \end{cases} \quad (2)$$

The following assumptions are understood:

1. Under any stress the lifetime of a test unit follows a Rayleigh distribution.
2. Testing is conducted at stresses X_1 and X_2, where $X_1 < X_2$.
3. The relationship between the parameter θ_i and the stress X_i is given by $Ln(\theta_i) = \beta_0 + \beta_1 X_i$, where β_0 and β_1 are unknown parameters to be determined from the test data.
4. The lifetimes of test units are independent and identically distributed.
5. All n units are initially placed on low stress X_1 and run until time τ when the stress is changed to high stress X_2. At X_2 testing continues until all remaining units fail.

Verification that the Rayleigh cumulative exposure model for step-stress is given by:

$$G(y) = \begin{cases} 1 - e^{\frac{-y^2}{\theta_1}} & , 0 < y < \tau \\ 1 - e^{\frac{-\left(y - \tau + \tau\sqrt{\frac{\theta_2}{\theta_1}}\right)^2}{\theta_2}} & , \tau < y < \infty \end{cases} \quad (3)$$

If $T = Y^2$, then the cumulative exposure model of T is given by:

$$G(t) = \begin{cases} 1 - e^{\frac{-t}{\theta_1}} & , 0 < t < \tau^2 \\ 1 - e^{\frac{-\left(\sqrt{t} - \tau + \tau\sqrt{\frac{\theta_2}{\theta_1}}\right)^2}{\theta_2}} & , \tau^2 < t < \infty \end{cases} \quad (4)$$

Note that $G(t)$ is not a step-stress exponential cumulative exposure model. For simplicity, let $\tau_1 = \tau^2$, so that the cumulative exposure model of T is given by:

$$G(t) = \begin{cases} 1 - e^{\frac{-t}{\theta_1}} & ,0 < t < \tau_1 \\ 1 - e^{\frac{-\left(\sqrt{t} - \sqrt{\tau_1} + \sqrt{\frac{\tau_1 \theta_2}{\theta_1}}\right)^2}{\theta_2}} & ,\tau_1 < t < \infty \end{cases} \quad (5)$$

and the corresponding probability density function of T is given by:

$$g(t) = \begin{cases} \frac{1}{\theta_1} e^{\frac{-t}{\theta_1}} & ,0 < t < \tau_1 \\ \frac{\left(\sqrt{t} - \sqrt{\tau_1} + \sqrt{\frac{\tau_1 \theta_2}{\theta_1}}\right)}{\theta_2 \sqrt{t}} e^{\frac{-\left(\sqrt{t} - \sqrt{\tau_1} + \sqrt{\frac{\tau_1 \theta_2}{\theta_1}}\right)^2}{\theta_2}} & ,\tau_1 < t < \infty \end{cases} \quad (6)$$

Methodology

Model Parameters Estimation

Let t_{ij}, $j = 1, 2, ..., n_i$, $i = 1, 2$ be the observed lifetime under low and high stress, where n_1 denotes the number of units failed at the low stress X_1 and n_2 denotes the number of units failed at the high stress X_2. The Likelihood function is given by:

$$L(t_{ij}, \beta_0, \beta_1) = \theta_1^{-n_1} \theta_2^{-n2} \; e^{-\frac{\sum_{j=1}^{n_1} t_{1j}}{\theta_1} - \frac{\sum_{j=1}^{n_2} \left(\sqrt{t_{2j}} - \sqrt{\tau_1} + \sqrt{\frac{\tau_1 \theta_2}{\theta_1}}\right)^2}{\theta_2}} * \prod_{j=1}^{n_2} \left(1 - \sqrt{\frac{\tau_1}{t_{2j}}} + \sqrt{\frac{\tau_1 \theta_2}{t_{2j} \theta_1}}\right) \quad (7)$$

where

$\theta_1 = e^{\beta_0 + \beta_1 X_1}$ and $\theta_2 = e^{\beta_0 + \beta_1 X_2}$.

The log likelihood function is given by:

$$LnL(t_{ij}, \beta_0, \beta_1) = - n_1 \ln(\theta_1) - n_2 \ln(\theta_2) + \sum_{j=1}^{n_2} Ln\left(1 - \sqrt{\frac{\tau_1}{t_{2j}}} + \sqrt{\frac{\tau_1 \theta_2}{t_{2j} \theta_1}}\right) - \frac{\sum_{j=1}^{n_1} t_{1j}}{\theta_1} - \frac{\sum_{j=1}^{n_2} \left(\sqrt{t_{2j}} - \sqrt{\tau_1} + \sqrt{\frac{\tau_1 \theta_2}{\theta_1}}\right)^2}{\theta_2} \quad (8)$$

The maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model parameters β_0 and β_1 can be obtained by solving numerically the following two equations:

$$\frac{\partial LnL(t_{ij}, \beta_0, \beta_1)}{\partial \beta_0} = 0 \quad (9)$$

$$\frac{\partial LnL(t_{ij}, \beta_0, \beta_1)}{\partial \beta_1} = 0 \quad (10)$$

In order to construct confidence intervals for the model parameters, the asymptotic normality of the maximum likelihood estimates are used. It is known that:

$$(\hat{\beta}_0, \hat{\beta}_1) \sim N\left((\beta_0, \beta_1), \hat{F}^{-1}\right)$$

where $\hat{F}^{-1}$ denotes the inverse of the observed Fisher information matrix $\hat{F}$. The observed Fisher information matrix $\hat{F}$ is obtained by evaluating the second and mixed partial derivatives of $LnL(t_{ij}, \beta_0, \beta_1)$ at the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, that is:

$$\hat{F} = \begin{bmatrix} \hat{F}_{11} & \hat{F}_{12} \\ \hat{F}_{21} & \hat{F}_{22} \end{bmatrix}$$

where

$$\hat{F}_{11} = -\frac{\partial^2 LnL(t_{ij}, \beta_0, \beta_1)}{\partial \beta_0^2}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1}$$

$$\hat{F}_{12} = \hat{F}_{21} = -\frac{\partial^2 LnL(t_{ij}, \beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1}$$

$$\hat{F}_{22} = -\frac{\partial^2 LnL(t_{ij}, \beta_0, \beta_1)}{\partial \beta_1^2}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1}$$

Thus, a $(1-\alpha)100$ % confidence interval for the model parameter $\beta_l, l = 0, 1$ is given by:

$$[\hat{\beta}_l - Z_{1-\alpha/2}\, S.E(\hat{\beta}_l), \quad \hat{\beta}_l + Z_{1-\alpha/2}\, S.E(\hat{\beta}_l)]$$

where $S.E(\hat{\beta}_l)$ denotes the standard error of the maximum likelihood estimates $\hat{\beta}_l$ which is the square root of the diagonal element of $\hat{F}^{-1}$, and $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentile of the standard normal distribution.

Note that an optimal test plan can be determined by minimizing with respect to the change time τ_1 the asymptotic variance at the design stress X_0. Thus, the numerical search method was used to find the value of τ_1^* that minimizes $(1 \quad X_0)\,\hat{F}^{-1}\,(1 \quad X_0)^T$, where $(1 \quad X_0)^T$ denotes the transpose of the vector $(1 \quad X_0)$. Thus the optimum time to change stress under the Rayleigh cumulative exposure model is $\tau^* = \sqrt{\tau_1^*}$.

Example

The data in Table 1 includes $n = (n_1 + n_2) = 30$ simulated observations from cumulative exposure model (5) defined above. The values used in this example are: $\beta_0 = 2$, $\beta_1 = 4$, $\tau_1^* = 28.4$.

Table 1: Simulated Data

Stress	Failure Times		
$X_1 = 0.2$	4.49977	7.17595	13.8035
	20.4499	1.89708	12.0347
	20.8349	23.9302	7.48286
	4.14598	0.101846	0.535875
	7.20451	12.7104	14.0179
	13.2136	23.3564	26.5207
	8.30107	1.3215	
$X_2 = 1$	262.761		645.625
	152.777		81.7587
	589.63		65.7081
	575.368		100.604
	168.515		281.587

The simulated data results show:

1. The values of the maximum likelihood estimates are $\hat{\beta}_0 = 2.45473$, $\hat{\beta}_1 = 4.10729$.
2. The inverse of the observed Fisher information matrix is $\hat{F}^{-1} = \begin{bmatrix} 0.119688 & -0.235227 \\ -0.235227 & 0.640747 \end{bmatrix}$
3. A 95% confidence interval for β_0 is [1.77665, 3.13281].
4. A 95% confidence interval for β_1 is [2.53838, 5.6762].

Simulation Study

A simulation study was conducted to investigate the performance of the maximum likelihood estimates, and the performance of the confidence interval based on the asymptotic normality of the maximum likelihood estimates. The criteria used for the evaluation of the performance of the maximum likelihood estimates were the bias and the mean squared error (MSE). For the confidence interval with confidence coefficient $(1-\alpha)$ the following were calculated:

1. The expected width (W): the average width of the simulated intervals.
2. Lower error rate (L): the fraction of intervals that fall entirely above the true parameter.

3. Upper error rate (U): the fraction of intervals that fall entirely below the true parameters.
4. Total error rate (T): the fraction of intervals that did not contain the true parameter value.

The indices of the simulation study were:

- n: total number of units placed on the test, n = 10, 40, 80, 100.
- X_1: low stress level, X_1 = 0.1, 0.2, 0.3, 0.5.
- X_2: high stress level, X_2 = 0.9, 1.0, 1.2, 1.3, 1.9.
- For $\beta_0 = 4$, $\beta_1 = 6$, α = 0.05 and for each combination of n, X_1 and X_2 2,000 samples were generated.

Results

Tables 2-5 show simulation results for parameter β_0, while Tables 6-9 show simulation results for parameter β_1.

Conclusion

Based on the simulation results the following conclusions are put forth. For the parameter β_0, the maximum likelihood estimate $\hat{\beta}_0$ has small values of bias and mean squared error, also as the sample size increases the value of the bias and the mean squared error decreases. The confidence interval for $\hat{\beta}_0$ had a small expected width value and the expected width decreases as the sample size increases. In terms of attainment of the coverage probability and the symmetry of lower and upper error rates, the intervals behave very well especially for large value of n. Also, from the results it appears that, for the same value of X_2, as the value of X_1 increases the values of expected width, bias and mean squared error also increase. Conversely, for the same value of X_1, as the value of X_2 increases the values of expected width, bias and mean squared error decrease. Thus, the recommendation is to use a small value of X_1 and a large value of X_2, and the same conclusions can be drawn for the parameter β_1.

Table 2: Simulation Results of the Parameter β_0 when n = 10

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	2.1703	0.0440	0.0000	0.0440	0.2576	0.2089
	0.2	3.0980	0.0415	0.0000	0.0415	0.2607	0.4146
	0.3	4.1653	0.0505	0.0015	0.0520	0.3094	0.8260
	0.5	7.9925	0.0540	0.0040	0.0580	0.5476	3.1766
1.0	0.1	1.9623	0.0380	0.0005	0.0385	0.2499	0.1951
	0.2	3.0753	0.0390	0.0005	0.0395	0.2477	0.3675
	0.3	3.8550	0.0525	0.0025	0.0550	0.3097	0.7143
	0.5	6.0328	0.0435	0.0050	0.0485	0.4283	2.1923
1.2	0.1	1.9296	0.0430	0.0000	0.0430	0.2543	0.1796
	0.2	2.6484	0.0410	0.0010	0.0420	0.2520	0.2965
	0.3	3.3296	0.0520	0.0010	0.0530	0.3024	0.5133
	0.5	4.9964	0.0540	0.0025	0.0565	0.3945	1.3313
1.3	0.1	1.8004	0.0350	0.0005	0.0355	0.2496	0.1637
	0.2	2.4806	0.0435	0.0010	0.0445	0.2625	0.2787
	0.3	3.1185	0.0395	0.0030	0.0425	0.2600	0.4204
	0.5	4.6224	0.0445	0.0025	0.0470	0.3078	1.1109
1.9	0.1	1.6907	0.0460	0.0005	0.0465	0.2815	0.1631
	0.2	2.1678	0.0390	0.0010	0.0400	0.2471	0.1990
	0.3	2.5498	0.0325	0.0010	0.0335	0.2292	0.2569
	0.5	3.5230	0.0465	0.0015	0.0480	0.2602	0.5334

Table 3: Simulation Results of the Parameter β_0 when n = 40

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	0.9862	0.0375	0.0090	0.0465	0.0439	0.0563
	0.2	1.2743	0.0425	0.0125	0.0550	0.0447	0.1110
	0.3	1.7032	0.0415	0.0100	0.0515	0.0609	0.1921
	0.5	3.1698	0.0390	0.0155	0.0545	0.1073	0.6790
1.0	0.1	1.0098	0.0345	0.0105	0.0450	0.0491	0.0511
	0.2	1.2215	0.0460	0.0100	0.0560	0.0585	0.0969
	0.3	1.5599	0.0390	0.0125	0.0515	0.0621	0.1616
	0.5	2.6856	0.0355	0.0140	0.0495	0.0758	0.4816
1.2	0.1	0.9224	0.0360	0.0045	0.0405	0.0468	0.0450
	0.2	1.1081	0.0375	0.0105	0.0480	0.0368	0.0793
	0.3	1.3694	0.0350	0.0075	0.0425	0.0489	0.1197
	0.5	2.1300	0.0385	0.0175	0.0560	0.0587	0.3000
1.3	0.1	0.8921	0.0310	0.0100	0.0410	0.0360	0.0419
	0.2	1.0858	0.0325	0.0065	0.0390	0.0451	0.0686
	0.3	1.3059	0.0350	0.0135	0.0485	0.0508	0.1135
	0.5	1.9469	0.0360	0.0155	0.0515	0.0732	0.2551
1.9	0.1	0.8267	0.0290	0.0070	0.0360	0.0476	0.0328
	0.2	1.0231	0.0270	0.0060	0.0330	0.0411	0.0512
	0.3	1.1547	0.0360	0.0130	0.0490	0.0443	0.0767
	0.5	1.4178	0.0345	0.0110	0.0455	0.0602	0.1319

Table 4: Simulation Results of the Parameter β_0 when n = 80

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	0.6529	0.0345	0.0210	0.0555	0.0100	0.0276
	0.2	0.8844	0.0345	0.0170	0.0515	0.0305	0.0504
	0.3	1.1884	0.0380	0.0145	0.0525	0.0313	0.0932
	0.5	2.2149	0.0405	0.0160	0.0565	0.0706	0.3394
1.0	0.1	0.6313	0.0295	0.0140	0.0435	0.0134	0.0234
	0.2	0.8338	0.0325	0.0105	0.0430	0.0314	0.0443
	0.3	1.0872	0.0355	0.0195	0.0550	0.0269	0.0809
	0.5	1.8748	0.0340	0.0145	0.0485	0.0577	0.2303
1.2	0.1	0.6049	0.0330	0.0160	0.0490	0.0133	0.0238
	0.2	0.7612	0.0400	0.0130	0.0530	0.0254	0.0379
	0.3	0.9534	0.0350	0.0180	0.0530	0.0278	0.0596
	0.5	1.4828	0.0335	0.0125	0.0460	0.0403	0.1381
1.3	0.1	0.5973	0.0250	0.0120	0.0370	0.0121	0.0220
	0.2	0.7351	0.0395	0.0180	0.0575	0.0274	0.0361
	0.3	0.9059	0.0320	0.0205	0.0525	0.0249	0.0560
	0.5	1.3628	0.0360	0.0210	0.0570	0.0324	0.1259
1.9	0.1	0.5604	0.0250	0.0120	0.0370	0.0174	0.0173
	0.2	0.6401	0.0380	0.0180	0.0560	0.0190	0.0281
	0.3	0.7435	0.0295	0.0165	0.0460	0.0161	0.0338
	0.5	0.9887	0.0310	0.0165	0.0475	0.0200	0.0631

Table 5: Simulation Results of the Parameter β_0 when n = 100

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	0.5793	0.0285	0.0150	0.0435	0.0140	0.0211
	0.2	0.7890	0.0320	0.0195	0.0515	0.0198	0.0414
	0.3	1.0585	0.0460	0.0170	0.0630	0.0309	0.0787
	0.5	1.9750	0.0290	0.0150	0.0440	0.0467	0.2473
1.0	0.1	0.5612	0.0295	0.0095	0.0390	0.0178	0.0193
	0.2	0.7432	0.0305	0.0160	0.0465	0.0201	0.0354
	0.3	0.9699	0.0335	0.0175	0.0510	0.0295	0.0621
	0.5	1.6748	0.0325	0.0205	0.0530	0.0284	0.1858
1.2	0.1	0.5354	0.0255	0.0125	0.0380	0.0141	0.0169
	0.2	0.6790	0.0355	0.0190	0.0545	0.0158	0.0313
	0.3	0.8494	0.0375	0.0150	0.0525	0.0192	0.0465
	0.5	1.3251	0.0325	0.0205	0.0530	0.0227	0.1167
1.3	0.1	0.5261	0.0260	0.0150	0.0410	0.0172	0.0171
	0.2	0.6556	0.0250	0.0200	0.0450	0.0154	0.0264
	0.3	0.8070	0.0355	0.0160	0.0515	0.0225	0.0420
	0.5	1.2151	0.0290	0.0220	0.0510	0.0164	0.0977
1.9	0.1	0.4892	0.0275	0.0170	0.0445	0.0151	0.0150
	0.2	0.5702	0.0280	0.0170	0.0450	0.0137	0.0205
	0.3	0.6626	0.0315	0.0195	0.0510	0.0167	0.0283
	0.5	0.8811	0.0330	0.0205	0.0535	0.0145	0.0518

Table 6: Simulation Results of the Parameter β_1 when n = 10

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	11.7520	0.0000	0.1050	0.1050	-1.1302	5.4630
	0.2	10.7495	0.0000	0.0920	0.0920	-0.9623	5.0870
	0.3	10.6093	0.0005	0.0835	0.0840	-0.8507	5.5120
	0.5	13.7981	0.0020	0.0765	0.0785	-0.9891	9.4010
1.0	0.1	9.5277	0.0000	0.1020	0.1020	-0.9831	4.4160
	0.2	10.6398	0.0005	0.0980	0.0985	-0.8012	4.0510
	0.3	9.6039	0.0005	0.0945	0.0950	-0.8252	4.5300
	0.5	9.9503	0.0025	0.0700	0.0725	-0.7903	6.2630
1.2	0.1	9.0816	0.0000	0.1130	0.1130	-0.8944	3.1220
	0.2	8.4417	0.0000	0.0945	0.0945	-0.7595	2.8450
	0.3	7.8835	0.0015	0.0985	0.1000	-0.7365	2.9810
	0.5	7.9438	0.0005	0.0795	0.0800	-0.6901	3.5090
1.3	0.1	7.6321	0.0000	0.0980	0.0980	-0.7888	2.5240
	0.2	7.5746	0.0005	0.1070	0.1075	-0.7344	2.5700
	0.3	7.1972	0.0000	0.0840	0.0840	-0.6107	2.2610
	0.5	7.2356	0.0000	0.0755	0.0755	-0.5331	2.8070
1.9	0.1	6.0578	0.0000	0.1165	0.1165	-0.6138	1.3170
	0.2	5.9001	0.0000	0.1085	0.1085	-0.5047	1.1770
	0.3	5.3165	0.0000	0.0920	0.0920	-0.4502	1.0890
	0.5	5.1264	0.0010	0.0880	0.0890	-0.3977	1.1430

Table 7: Simulation Results of the Parameter β_1 when n = 40

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	5.4639	0.0035	0.0860	0.0895	-0.4575	2.0540
	0.2	4.0934	0.0050	0.0585	0.0635	-0.2193	1.2430
	0.3	4.0963	0.0080	0.0590	0.0670	-0.2089	1.2220
	0.5	5.2896	0.0110	0.0435	0.0545	-0.2028	1.9090
1.0	0.1	5.6772	0.0030	0.0920	0.0950	-0.4462	1.8450
	0.2	3.8432	0.0045	0.0620	0.0665	-0.2485	1.0380
	0.3	3.6264	0.0045	0.0545	0.0590	-0.1817	0.9110
	0.5	4.3539	0.0120	0.0485	0.0605	-0.1580	1.2890
1.2	0.1	4.8035	0.0020	0.0875	0.0895	-0.4210	1.4880
	0.2	3.3155	0.0065	0.0595	0.0660	-0.1993	0.8150
	0.3	3.0358	0.0055	0.0600	0.0655	-0.1688	0.6660
	0.5	3.2786	0.0105	0.0505	0.0610	-0.1244	0.7370
1.3	0.1	4.4699	0.0020	0.0700	0.0720	-0.3210	1.1760
	0.2	3.2118	0.0030	0.0710	0.0740	-0.2396	0.7610
	0.3	2.8369	0.0060	0.0595	0.0655	-0.1685	0.5970
	0.5	2.9173	0.0105	0.0490	0.0595	-0.1386	0.5930
1.9	0.1	3.6709	0.0010	0.0745	0.0755	-0.2878	0.7030
	0.2	2.9251	0.0020	0.0625	0.0645	-0.1974	0.4530
	0.3	2.3678	0.0050	0.0745	0.0795	-0.1508	0.3530
	0.5	1.9082	0.0055	0.0580	0.0635	-0.1121	0.2710

Table 8: Simulation Results of the Parameter β_1 when n = 80

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	3.3864	0.0045	0.0555	0.0600	-0.1634	0.8710
	0.2	2.8042	0.0110	0.0420	0.0530	-0.1308	0.5250
	0.3	2.8413	0.0140	0.0480	0.0620	-0.1052	0.5510
	0.5	3.6913	0.0115	0.0470	0.0585	-0.1321	0.9460
1.0	0.1	3.1606	0.0025	0.0580	0.0605	-0.1898	0.7530
	0.2	2.5651	0.0080	0.0480	0.0560	-0.1370	0.4590
	0.3	2.5175	0.0125	0.0450	0.0575	-0.0806	0.4390
	0.5	3.0276	0.0125	0.0415	0.0540	-0.1065	0.6020
1.2	0.1	2.8929	0.0080	0.0670	0.0750	-0.1713	0.6230
	0.2	2.2143	0.0090	0.0595	0.0685	-0.1175	0.3520
	0.3	2.0923	0.0110	0.0460	0.0570	-0.0861	0.2960
	0.5	2.2683	0.0115	0.0325	0.0440	-0.0689	0.3200
1.3	0.1	2.8233	0.0035	0.0675	0.0710	-0.1744	0.5990
	0.2	2.0904	0.0070	0.0550	0.0620	-0.1297	0.3180
	0.3	1.9392	0.0160	0.0475	0.0635	-0.0719	0.2680
	0.5	2.0396	0.0140	0.0520	0.0660	-0.0658	0.2840
1.9	0.1	2.3871	0.0030	0.0765	0.0795	-0.1722	0.4090
	0.2	1.6190	0.0065	0.0720	0.0785	-0.1069	0.2040
	0.3	1.4227	0.0085	0.0435	0.0520	-0.0585	0.1350
	0.5	1.3256	0.0120	0.0350	0.0470	-0.0468	0.1130

Table 9: Simulation Results of the Parameter β_1 when n = 100

X_2	X_1	W	L	U	T	Bias	MSE
0.9	0.1	2.9543	0.0100	0.0475	0.0575	-0.1281	0.6250
	0.2	2.4986	0.0140	0.0440	0.0580	-0.0878	0.4090
	0.3	2.5221	0.0140	0.0530	0.0670	-0.0830	0.4470
	0.5	3.2879	0.0120	0.0315	0.0435	-0.0776	0.6800
1.0	0.1	2.7718	0.0050	0.0505	0.0555	-0.1569	0.5480
	0.2	2.2803	0.0120	0.0450	0.0570	-0.0832	0.3500
	0.3	2.2438	0.0120	0.0430	0.0550	-0.0898	0.3420
	0.5	2.7064	0.0180	0.0345	0.0525	-0.0587	0.4950
1.2	0.1	2.5031	0.0070	0.0540	0.0610	-0.1534	0.4590
	0.2	1.9714	0.0080	0.0480	0.0560	-0.0793	0.2750
	0.3	1.8585	0.0115	0.0465	0.0580	-0.0556	0.2260
	0.5	2.0282	0.0125	0.0385	0.0510	-0.0463	0.2790
1.3	0.1	2.4068	0.0035	0.0670	0.0705	-0.1662	0.4800
	0.2	1.8635	0.0105	0.0395	0.0500	-0.0847	0.2290
	0.3	1.7225	0.0090	0.0440	0.0530	-0.0654	0.1960
	0.5	1.8186	0.0190	0.0355	0.0545	-0.0412	0.2250
1.9	0.1	1.9923	0.0035	0.0735	0.0770	-0.1514	0.3130
	0.2	1.4353	0.0075	0.0575	0.0650	-0.0729	0.1530
	0.3	1.2630	0.0085	0.0430	0.0515	-0.0534	0.1110
	0.5	1.1779	0.0115	0.0405	0.0520	-0.0352	0.0920

References

Al-Haj Ebrahem, M., & Al Masri, A. (2007(a)). Optimum simple step-stress plan for log-logistic cumulative exposure model. *Metron Iternational Journal of Statistics*, LXV(1), 23-34.

Al-Haj Ebrahem, M., & Al Masri, A. (2007(b)). Optimum simple step-stress plan for log-logistic distribution under time censoring. *Abhath Al-Yarmouk: Basic Sci. and Eng.*, 16, 319-327.

Alhadeed, A., & Yang, S. (2005). Optimal simple step-stress plan for cumulative exposure model using log-normal distribution. *IEEE Transactions on Reliability*, *54*, 64-68.

Bai, D., Kim, M., & Lee, S. (1989). Optimum simple step-stress accelerated life tests with censoring. *IEEE Transactions on Reliability*, 38, 528-532.

Chung, S., & Bai, D. (1998). Optimal designs of simple step-stress accelerated life tests for lognormal lifetime distributions. *International Journal of Reliability, Quality and Safety Engineering*, *5*, 315-336.

Cohen, A., & Whitten, B. (1988). *Parameter estimation in reliability and life span models*. New York, Marcel Dekker.

Gouno, E. (2001). An inference method for temperature step-stress accelerated life testing. *Quality and Reliability Engineering International*, *17*, 11-18.

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions: Volume 1*. New York: Wiley.

Khamis, I., & Higgins, J. (1996). Optimum 3-step step-stress tests. *IEEE Transactions on Reliability*, *45*, 341-345.

Khamis, I., & Higgins, J. (1996). An alternative to the Weibull step-stress model. *Proceedings of the American Statistical Association*, *Chicago*, 123-127.

Khamis, I., & Higgins, J. (1998). A new model for step-stress testing. *IEEE Transactions on Reliability*, *47*, 131-134.

Miller, R., & Nelson, B. (1983). Optimum simple step-stress plans for accelerated life testing. *IEEE Transactions on Reliability*, *32*, 59-65.

Nelson, W. (1980). Accelerated life testing step-stress models and data analysis. *IEEE Transactions on Reliability*, *29*, 103-108.

Nelson, W. (1990). *Accelerated testing, statistical models, test plans, and data analysis*. New York: Wiley.

Schmoyer, R. (1991). Nonparametric analysis for two-level single-stress accelerated life tests. *Technometrics*, *33*, 175-186.

Shaked, M., & Singurwalla, N. (1983). Inference for step-stress accelerated life tests. *Journal of Statistical Planning and Inference*, *7*, 295-306.

Xiong, C. (1998). Inferences on a simple step-stress model with type-II censored exponential data. *IEEE Transactions on Reliability*, *47*, 142-146.

Xiong, C., & Milliken, G. (2002). Prediction for exponential lifetimes based on step-stress testing. *Communications in Statistics-Simulation*, *31*, 539-556.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 488-504

1538 – 9472/09/$95.00

Estimating Model Complexity of Feed-Forward Neural Networks

Douglas Landsittel
University of Pittsburgh

In a previous simulation study, the complexity of neural networks for limited cases of binary and normally-distributed variables based the null distribution of the likelihood ratio statistic and the corresponding chi-square distribution was characterized. This study expands on those results and presents a more general formulation for calculating degrees of freedom.

Key words: Degrees of freedom, null distribution, chi-square distribution.

Introduction

Feed-forward neural networks are commonly utilized as a statistical tool for classification and prediction of high-dimensional and/or potentially highly non-linear data. Their popularity stems from an implicitly non-linear and flexible model structure, which does not require explicit specification of interactions or other non-linear terms, and can universally approximate any function (Ripley, 1996). In cases where epidemiologic data or the underlying theory of the specific problem suggest a complex association, but the exact nature of such associations is not well understood, neural networks represent a more flexible methodology for potentially modeling such associations. One significantly negative consequence of this implicit non-linearity and flexible model structure, however, is the resulting inability to quantify model complexity. The typical approach of counting model terms does not provide a rationale basis for quantifying the effective model dimension because the model parameters are inherently correlated to varying degrees.

Douglas Landsittel is an Associate Professor of Medicine and Clinical and Translational Science in the Division of General Internal Medicine. Email: dpl12@pitt.edu.

Previous work has sought to quantify model degrees of freedom for other nonlinear or nonparametric models through use of the hat matrix. For scatterplot smoothers, local regression, and other nonparametric models, Hastie and Tibshirani (1990) and others directly calculated the trace of the hat matrix to estimate degrees of freedom. In cases where the hat matrix cannot be explicitly specified, such as more complex models or model selection procedures, Ye (1998) proposes the generalized degrees of freedom, which estimates the diagonal terms based on the sensitivity of fitted values to changes in observed response values. To address random effects, hierarchical models, and other regression methods, Hodges and Sargent (2001) extended degrees of freedom using a re-parameterization of the trace of the hat matrix and subsequent linear model theory.

Other publications have specifically addressed the issue of model complexity for neural networks. For instance, Moody (1992) calculated the effective number of model parameters based on approximating the test set error as a function of the training set error plus model complexity. A number of other articles (Liu, 1995; Amari & Murata, 1993; Murata, Yoshizawa, & Amari, 1991) have presented theorems to quantify model complexity, but, without a framework for practically applying such methods, none have been utilized in practice. Others have taken a more computational approach (as summarized by Ripley, 1996; and Tetko, Villa, & Livingstone, 1996) using methods such as cross-validation, eliminating variables based on small (absolute)

parameter values, or eliminating variables with a small effect on predicted values (i.e. sensitivity methods). Bayesian approaches have also been proposed (Ripley, 1996; Paige & Butler, 2001) for model selection with neural networks. Despite the noted advances, implementation of such methods has been limited by either computational issues, dependence on the specified test set, or lack of distributional theory. As a result, there are no established procedures for variable selection or determination of the optimal network structure (e.g. the number of hidden units) with neural networks.

Previously, a simulation study was conducted to investigate the distribution of the likelihood ratio statistic with neural networks. In the present study, simulations are conducted to empirically describe the distribution of the likelihood ratio statistic under the null assumption of the intercept model (versus the alternative of at least one non-zero covariate parameter). All simulations are conducted with a single binary response; in contrast, the previously cited literature primarily focuses on continuous outcomes. In cases where the likelihood ratio can be adequately approximated by a chi-square distribution, the degrees of freedom can be used to quantify neural network model complexity under the null. Derivation of the test statistic null distribution is pursued through simulation approaches, rather than theoretical derivations, because of the complexity of the network response function and the lack of maximum likelihood or other globally optimal estimation.

The two main objectives of this simulation study are to: (1) verify that the chi-square distribution provides an adequate approximation to the empirical test statistic distribution in a limited number of simulated cases, and (2) quantify how the distribution, number of covariates and the number of hidden units affects model degrees of freedom. Adequacy of the chi-square approximation will be judged by how close the α-level based on the simulation distribution (i.e., the percent of the test statistic distribution greater than the corresponding chi-square quantile) is to various percentiles of the chi-square distribution. The variance, which should be approximately twice the mean under a chi-square distribution, is also displayed for each simulation condition.

Methodology

The Feed-Forward Neural Network Model

This study focuses strictly on a single Bernoulli outcome, such as presence or absence of disease. All neural network models utilized a feed-forward structure (Ripley, 1996) with a single hidden layer so that

$$\hat{y}_k = f(v_0 + \sum_{j=1}^{H} v_j f\{w_{jo} + \sum_{i=1}^{p} w_{ji} x_{ik}\}), \quad (1)$$

where $\hat{y}$ is the predicted value for the k^{th} observation with covariate values $\boldsymbol{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})$. The function $f(x)$ is the logistic function, $1/(1+e^{-x})$, and each logistic function, $f\{w_{jo} + \sum_{i=1}^{p} w_{ji} x_{ik}\}$, is referred to as the j^{th} hidden unit. The response function of the neural network model can thus be viewed as a logistic of these hidden unit values. In terms of further terminology, the parameters $v_0, v_1, ..., v_H$ are referred to as the connections between the hidden and output layer and each set of other parameters, $w_{j1}, w_{j2}, ..., w_{jp}$, are referred to as the connections between the inputs and hidden units, where there are *p* covariate values specific each of the *p* hidden units. This described model structure often leads to categorization of neural networks as a black box technique. None of the parameter values directly correspond to any specific main effect or interaction. Further, the degree of non-linearity cannot be explicitly determined from the number of hidden units or any easily characterized aspect of the model.

The optimal model coefficients were calculated via back-propagation (Rumelhart, et al., 1995) and the nnet routine in S-Plus (Venables & Ripley, 1997), which iteratively updates weights using a gradient descent-based algorithm. For a Bernoulli outcome, optimization is based on the minimization of the deviance (*D)*,

$$D = -2\sum_{(k=1)}^{n}[y_k log(\hat{y}_k) + (1 - y_k) log(1 - \hat{y}_k)] \quad (2)$$

with a penalty term for the sum of the squared weights (referred to as weight decay). Weight decay, represented by λ in Equation 3, is commonly utilized to improve optimization and generalization of the resulting model by minimizing the penalized likelihood (*PL*)

$$PL = D + \lambda \sum_{(j=1)}^{H}\left[v_j^2 + w_{ji}^2\right] \quad (3)$$

For this study, λ = 0.01 was utilized in all simulations based on previous experience and recommendations by Ripley (1996) established on Bayesian arguments and the range of the logistic function.

Quantifying Model Complexity through Simulations

All simulations utilized a feed-forward neural network with one hidden layer and a single binary outcome with a varying number of predictor variables. All variables were randomly generated (via S-Plus), with the predictor variables being simulated independently of the binary outcome, as to generate results under the null hypothesis of no association. Separate sets of simulations were conducted for a variety of conditions, including binary versus continuous predictors, a varying number of predictor variables, and a varying number of hidden units. For each condition, 500 data sets were each simulated, each with 2,000 observations (to approximate asymptotic results).

To quantify model complexity of the given neural network under the given set of conditions, the likelihood ratio statistic for model independence was calculated, which serves to quantify the model complexity under the null of no association between outcome and predictors. The simulations result in a distribution of likelihood ratios which should follow a chi-square distribution with the mean equal to the degrees of freedom. The mean of that distribution can then be used to quantify model complexity under the null. However, correspondence to a given chi-square distribution must be verified. In the absence of any current theoretical justification for the expected distribution, percentiles of the chi-square distribution were compared to the corresponding α-levels of the simulated distribution (of likelihood ratios). Simulated α-levels ($\alpha_q^{(S)}$) were then defined as the percentage of simulated values greater than q^{th} percentile of the corresponding chi-square distribution. For instance, the nominal α-level for the simulated distribution is given by

$$\alpha_{0.05}^{s} = P\left(LR \geq \chi_{0.05}^{2}\left(LR\right)\right) \quad (4)$$

where *LR* represents the likelihood ratio. Simulated α-levels are then compared to the chi-square percentiles at significance levels of 0.75, 0.50, 0.25, 0.10, and 0.05. Q-Q plots are also presented to quantify agreement with the appropriate chi-square distribution.

Methods for Estimating Model Degrees of Freedom

After verifying the expected correspondence to a chi-square distribution for a given set of conditions, a new method was utilized to estimate the degrees of freedom for other sets of conditions. Since these methods vary substantially for binary versus continuous predictors, the corresponding methods are first presented separately, after their respective simulation results, and then merged into a single approach. The actual methodology is presented within the results section since these methods are intuitively motivated by the simulation results, and are thus easier to understand within that context.

Results

Simulation Results for Binary Input Variables

Results presented in Table 1 were generated using independently distributed binary inputs. All neural network models were fit using a weight decay of 0.01; for each result pertaining to binary inputs, the maximum number of terms, including all main effects and interactions, for *k* inputs equals $2^k - 1$. The number of model parameters for a model with *h* hidden units equals $h(k + 1) + (h + 1)$.

Table 1: Likelihood Ratio Statistic for All Binary Inputs

Inputs (Max # Terms)	Hidden Units	#Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
2 (3)	2	9	2.86	5.81	0.75	0.52	0.26	0.06	0.03
	5	21	3.04	5.73	0.74	0.50	0.26	0.09	0.04
	10	41	3.15	5.98	0.76	0.51	0.25	0.10	0.04
3 (7)	2	11	7.04	15.90	0.75	0.49	0.22	0.09	0.04
	5	26	6.93	12.37	0.74	0.51	0.25	0.11	0.07
	10	51	7.24	13.97	0.75	0.50	0.26	0.10	0.06
4 (15)	2	13	11.94	22.05	0.74	0.50	0.25	0.11	0.08
	5	31	14.87	28.99	0.76	0.50	0.26	0.09	0.06
	10	61	14.96	31.33	0.76	0.50	0.23	0.08	0.04
5 (31)	2	15	18.36	31.03	0.75	0.50	0.26	0.13	0.08
	5	36	30.25	62.22	0.75	0.48	0.25	0.10	0.05
	10	71	31.82	69.57	0.74	0.50	0.22	0.09	0.06
6 (63)	2	17	25.07	44.05	0.71	0.49	0.28	0.14	0.07
	5	41	50.63	108.5	0.76	0.51	0.23	0.09	0.04
	10	81	63.70	147.5	0.76	0.50	0.24	0.08	0.03
7 (127)	2	19	30.92	57.98	0.74	0.50	0.26	0.10	0.05
	5	46	69.93	138.4	0.75	0.54	0.24	0.10	0.05
	10	91	117.3	245.6	0.75	0.50	0.25	0.10	0.05
8 (255)	2	21	38.75	77.43	0.74	0.51	0.25	0.08	0.04
	5	51	88.95	161.2	0.73	0.49	0.27	0.13	0.06
	10	101	168.3	318.0	0.74	0.50	0.27	0.11	0.05
9 (511)	2	23	45.76	110.9	0.79	0.51	0.20	0.06	0.02
	5	56	107.7	202.9	0.75	0.54	0.25	0.10	0.05
	10	111	214.4	394.9	0.74	0.50	0.24	0.11	0.06
10 (1023)	2	25	51.76	117.9	0.77	0.51	0.22	0.07	0.03
	5	61	126.1	248.5	0.74	0.51	0.24	0.10	0.05
	10	121	257.5	546.5	0.75	0.48	0.25	0.10	0.05
Mean Simulated α-levels					0.75	0.50	0.25	0.10	0.05

Table 1 shows the simulated distribution of the likelihood ratio test for independence is closely followed by a chi-square distribution. In a large percentage of the cases, all of the simulated α-levels were within 1-3% of the expected percentiles. No systematic differences were evident in the results. Figures 1a and 1b illustrate two examples where: (1) the simulated distribution varied a few percent from the expected percentiles (2 inputs and 2 hidden units), and (2) the simulated distribution fell extremely close to the corresponding chi-square distribution (7 inputs and 10 hidden units). Both figures show noticeable variability at the upper end of the distribution; however, it should be noted that these few points are primarily within only the top 1% of the distribution, and thus have little effect on most of the resulting significance levels.

Figure 1a: Example Q-Q plot (with 2 Binary Inputs, 2 HUs and 0.01 WD)
Illustrating Greater Variability from the Expected Chi-square Distribution

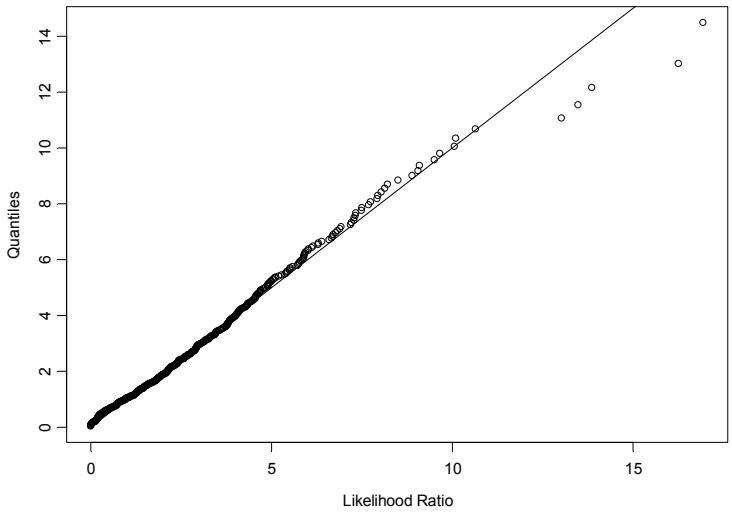


Figure 1b: Example Q-Q plot (with 7 Binary Inputs, 10 HUs and 0.01 WD)
Illustrating a Close Correspondence with the Expected Chi-square Distribution

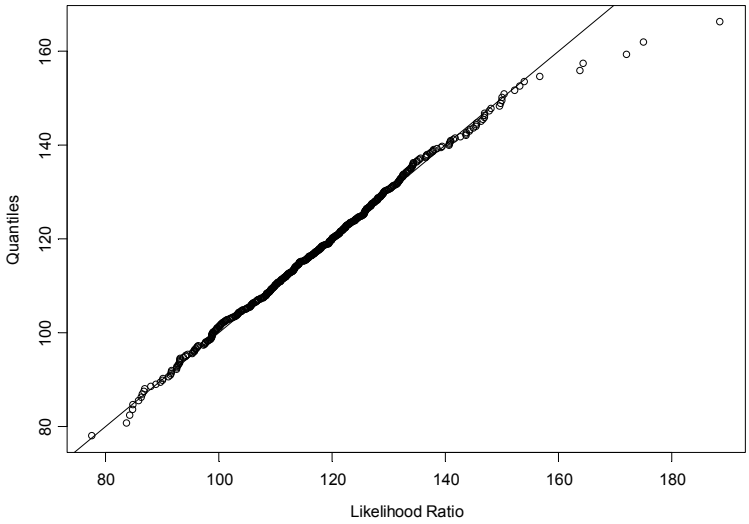


Table 2 shows additional simulation results for other cases (with at most 10 inputs) where the number of model parameters (p) is less than the maximum number of terms (m). Results again showed that the simulated α-levels were very close the expected percentiles. Q-Q plots for these cases (not shown here) resulted in similar findings as displayed in Figures 1a and 1b. No additional simulations are shown here for two, three or four inputs because these cases all corresponded to models where $p > m$, and will therefore produce a known degrees of freedom (3, 7, and 15 for 2, 3 and 4 inputs, respectively). Simulations were conducted, but not shown here, to verify that such results would hold for 4 hidden units (since p was only slightly greater than m in the case of 3 hidden units); results verified the expected finding (of 15 degrees of freedom). Other combinations leading to a known degrees of freedom (with $p > m$) were also excluded from the table (including 5 inputs with 6-9 hidden units and 6 inputs with 8-9 hidden units).

Estimating Degrees of Freedom for Binary Input Variables

The above results indicate that the model degrees of freedom for strictly binary inputs appear to intuitively depend on two factors:

(1) the maximum number of possible main effects and interactions = $(2^k - 1)$, and
(2) the number of model parameters = $h(k + 1) + (h + 1)$.

In cases where the number of model parameters is sufficient to fit all main effects and interactions, the degrees of freedom is equal to that maximum number of terms. For example, regardless of the number of hidden units, the degrees of freedom (*df*) are approximately 3.0 for two binary inputs and approximately 7.0 for three binary inputs. For four binary inputs, two hidden units (and subsequently 13 parameters) are insufficient to fit all 15 terms and result in approximately 12 *df*.

In such cases, where the number of model parameters is less than the maximum number of terms, the *df* is generally in between (or at least very close to) the number of model parameters (p) and the maximum number of terms (m). Exactly where the *df* falls depends on how close the number of model parameters is to the maximum number of terms. In general, the ratio of degrees of freedom by number of model parameters may be expressed as a function of $m - p$. To produce a linear relationship, it is more convenient (with binary inputs) to express df/p as a function of $\log_2(m\text{-}p)$. The simulated degrees of freedom from Table 1 was used to derive a relationship, and Figure 2 shows a plot of the simulated data from Table 1 (with 2, 5 or 10 hidden units) overlaid with the linear regression line

$$df/p = 0.6643 + 0.1429 \times log_2(m - p). \quad (5)$$

Figure 2 shows a general trend between the difference in $m - p$ and the degrees of freedom (divided by the number of parameters), but also illustrates some variability between the simulated values and the subsequent estimates. To evaluate the significance of these discrepancies, the estimated *df* were compared to the simulated distribution of the likelihood ratio statistic (for model independence). Results are shown in Table 3.

Results indicate that the estimated *df* usually approximates the simulated value within an absolute error of a few percent. For example, most of the conditions (11 of 16) result in a 5% significance level between 0.03 and 0.07; the largest discrepancy is an absolute difference of 0.04 from the true 5% level. The 10% significance level corresponds to somewhat larger errors, with the estimated p-values as high as 0.17 and as low as 0.02. The 75th, 50th and 25th percentiles showed similar findings with occasionally substantial discrepancies.

The above rule for estimating the *df*, based on the previously fit linear regression of df/p as a function of $log_2(m - p)$, was also evaluated with respect to its adequacy to predict model complexity for new cases (with 3, 4, or 6-9 hidden units). Figure 3 shows a plot of these additional simulated data overlaid with the linear same regression line $df/p = 0.6643 + 0.1429 \cdot log_2(m - p)$.

Figure 3 shows the trend between the difference in $m - p$ and the *df* (divided by the number of parameters), but again illustrates variability between the simulated values and the

Table 2: Additional Simulated Likelihood Ratio Statistics with All Binary Inputs

Inputs (Max # Terms)	Hidden Units	#Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
5 (31)	3	22	23.84	47.66	0.73	0.50	0.26	0.10	0.06
	4	29	28.69	54.80	0.74	0.50	0.26	0.12	0.06
6 (63)	3	25	34.57	67.28	0.74	0.49	0.25	0.10	0.06
	4	33	42.81	96.51	0.77	0.51	0.24	0.08	0.03
	6	49	55.86	109.8	0.74	0.50	0.26	0.09	0.05
	7	57	59.60	110.0	0.77	0.49	0.26	0.11	0.05
7 (127)	3	28	45.93	92.58	0.75	0.50	0.23	0.10	0.05
	4	37	58.06	111.6	0.75	0.50	0.24	0.10	0.05
	6	55	82.27	148.3	0.75	0.49	0.26	0.11	0.06
	7	64	92.84	175.6	0.74	0.51	0.27	0.10	0.05
	8	73	102.5	189.5	0.75	0.51	0.25	0.09	0.06
	9	82	111.7	224.0	0.76	0.51	0.24	0.10	0.06
8 (255)	3	31	54.90	101.0	0.75	0.50	0.26	0.11	0.07
	4	41	73.02	148.2	0.75	0.52	0.23	0.08	0.04
	6	61	107.8	223.0	0.75	0.49	0.24	0.09	0.05
	7	71	124.8	258.3	0.76	0.50	0.25	0.10	0.03
	8	81	139.7	238.2	0.71	0.52	0.28	0.12	0.06
	9	91	155.0	268.0	0.73	0.52	0.24	0.13	0.08
9 (511)	3	34	65.13	135.0	0.77	0.50	0.24	0.10	0.05
	4	45	87.02	179.6	0.76	0.51	0.25	0.09	0.04
	6	67	131.4	228.8	0.73	0.51	0.27	0.10	0.06
	7	78	152.3	286.6	0.74	0.50	0.25	0.10	0.06
	8	89	171.8	338.5	0.74	0.51	0.26	0.11	0.05
	9	100	194.7	303.6	0.72	0.50	0.27	0.14	0.08
10 (1023)	3	37	75.5	163.5	0.76	0.51	0.25	0.08	0.03
	4	49	100.9	190.8	0.73	0.52	0.26	0.10	0.05
	6	73	152.7	297.1	0.77	0.50	0.24	0.10	0.05
	7	85	178.5	341.9	0.74	0.51	0.24	0.09	0.06
	8	97	204.8	430.0	0.77	0.51	0.24	0.08	0.04
	9	109	230.0	425.7	0.74	0.52	0.25	0.10	0.06
Mean Simulated α-levels					0.75	0.51	0.25	0.10	0.05

Figure 2: Plot of the Degrees of Freedom for Binary Inputs (2, 5, and 10 Hidden Units) as a Function of the Difference between Maximum Number of Terms and Number of Parameters

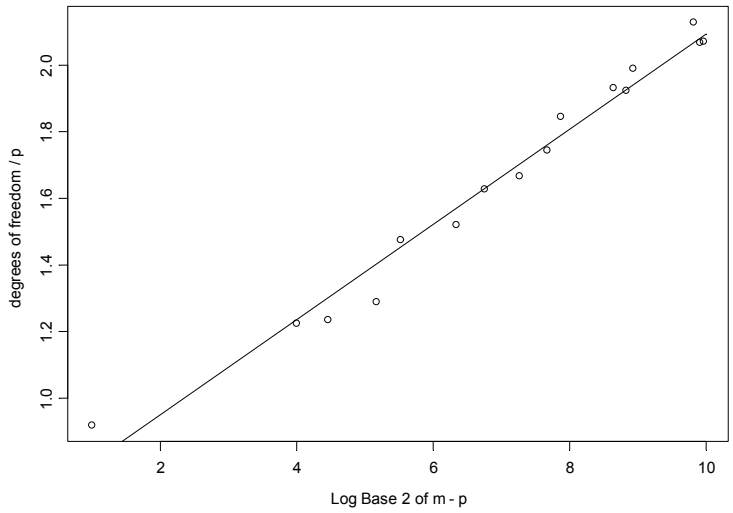
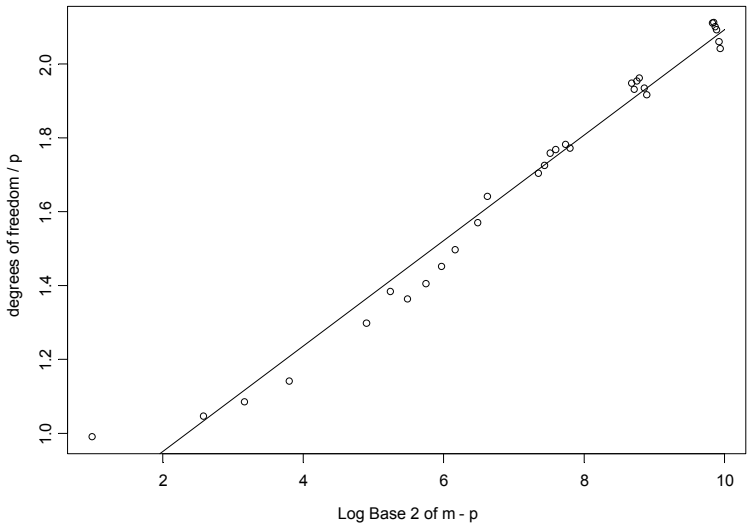


Table 3: Comparison of Estimated to Simulated Degrees of Freedom with All Binary Inputs

Max # of Terms (Inputs)	# of Parameters	Hidden Units	Simulated *df*	Estimated *df*	Simulated α-levels using the Estimated *df*				
					0.75	0.50	0.25	0.10	0.05
15 (4)	13	2	11.94	10.49	0.66	0.41	0.19	0.07	0.05
31 (5)	15	2	18.36	18.54	0.74	0.49	0.25	0.12	0.07
63 (6)	17	2	25.07	24.71	0.73	0.51	0.30	0.15	0.08
	41	5	50.63	53.36	0.67	0.40	0.16	0.05	0.02
127 (7)	19	2	30.92	30.96	0.74	0.50	0.26	0.10	0.05
	46	5	69.93	72.23	0.69	0.46	0.18	0.07	0.03
	91	10	117.3	127.7	0.50	0.25	0.09	0.02	0.01
255 (8)	21	2	38.75	37.57	0.78	0.56	0.30	0.11	0.05
	51	5	88.95	89.80	0.71	0.47	0.25	0.11	0.06
	101	10	168.3	172.0	0.67	0.42	0.21	0.07	0.03
511 (9)	23	2	45.76	44.63	0.82	0.56	0.24	0.08	0.03
	56	5	107.7	107.9	0.75	0.53	0.24	0.10	0.05
	111	10	214.4	210.8	0.80	0.57	0.30	0.15	0.08
1023 (10)	25	2	51.76	52.21	0.76	0.49	0.21	0.07	0.03
	61	5	126.1	126.9	0.72	0.49	0.23	0.09	0.05
	121	10	257.5	250.2	0.84	0.61	0.36	0.17	0.09
Mean Estimated α-levels					0.72	0.57	0.24	0.10	0.05

Figure 3: Plot of the Degrees of Freedom for Binary Inputs (3, 4, and 6-9 Hidden Units) as a Function of the Difference between Maximum Number of Terms and Number of Parameters



subsequent estimates. Further, results do not show any systematic difference from the previous set of findings (as graphed in Figure 2). To evaluate the significance of these discrepancies, the estimated *df* were compared to the simulated distribution of the likelihood ratio statistic (for model independence). Results are shown in Table 4.

Results again indicate that the estimated degrees of freedom usually approximated the simulated value within an absolute error of a few percent. For example, most of the conditions (20 of 30) resulted in a 5% significance level between 0.03 and 0.07; with two exceptions, the largest discrepancy is an absolute difference of 0.04 from the true 5% level. The 10% significance level, however, again corresponds to somewhat larger errors, with the estimated p-values being as high as 0.34 and as low as 0.04; most results (19 of 30), however, were between 0.07 and 0.13. The 75^{th}, 50^{th} and 25^{th} percentiles showed similar findings with occasionally higher discrepancies.

The above results identify some complications and discrepancies that arise when using this method to estimate the model *df* for strictly binary inputs. First, the subsequent simulations show only a fair degree of correspondence between the predicted and simulated *df*. The majority of conditions led to percentiles within an absolute difference of a few percent, but other conditions led to more substantial discrepancies. Secondly, the established rules under this method led to some logical inconsistencies in the predicted *df*. For example, with 5 inputs, the predicted *df* for 3 hidden units (24.58) is actually larger than that predicted for 4 hidden units (23.41). This apparent contradiction arises from the fact that the *df* are essentially predicted by scaling the number of parameters by a function of the difference between the maximum number of terms and the number of model parameters. While this approach has some intuitive appeal - and generally leads to an increase in the degrees of freedom as the number of hidden units increases (for a given number of input variables) - no guarantee exists that this pattern will hold universally.

Due to this, some corrections are therefore needed for predicting the model *df* in these scenarios. To do so, when a decrease is observed with an increase in hidden units, it is possible to simply take the average of the previous result with the next number of hidden units. For example, for the case of 5 inputs with 4 hidden units, the previous result (24.58 for 3 hidden units) would be averaged with the next result (31 for 5 hidden units) to obtain 27.79, which is much closer to the simulated result of

Table 4: Comparison of Estimated to Simulated Degrees of Freedom with Binary Inputs

Max #of Terms (Inputs)	# of Parameters	(Hidden Untis)	Simulated *df*	Estimated *df*	Simulated α-levels using the Estimated *df*				
					0.75	0.50	0.25	0.10	0.05
31 (5)	22	3	23.84	24.58	0.69	0.46	0.22	0.08	0.05
	29	4	28.69	23.41	0.91	0.77	0.55	0.34	0.21
63 (6)	25	3	34.57	35.36	0.71	0.45	0.22	0.08	0.05
	33	4	42.81	45.06	0.69	0.42	0.17	0.05	0.02
	49	6	55.86	59.21	0.63	0.38	0.17	0.05	0.03
	57	7	59.60	58.92	0.79	0.52	0.28	0.13	0.06
127 (7)	28	3	45.93	45.13	0.78	0.53	0.26	0.11	0.06
	37	4	58.06	58.90	0.73	0.47	0.22	0.09	0.04
	55	6	82.27	85.03	0.68	0.41	0.20	0.07	0.04
	64	7	92.84	97.18	0.63	0.38	0.17	0.05	0.02
	73	8	102.5	108.5	0.61	0.35	0.14	0.04	0.02
	82	9	111.7	118.8	0.59	0.33	0.12	0.04	0.02
255 (8)	31	3	54.90	55.18	0.74	0.49	0.25	0.11	0.06
	41	4	73.02	72.59	0.76	0.54	0.24	0.09	0.05
	61	6	107.8	106.77	0.78	0.52	0.26	0.11	0.06
	71	7	124.8	123.50	0.78	0.53	0.28	0.11	0.04
	81	8	139.7	139.96	0.70	0.51	0.27	0.12	0.06
	91	9	155.0	156.13	0.71	0.50	0.23	0.11	0.07
511 (9)	34	3	65.13	65.82	0.75	0.48	0.22	0.09	0.04
	45	4	87.02	86.89	0.76	0.51	0.26	0.09	0.04
	67	6	131.4	128.7	0.79	0.58	0.33	0.14	0.09
	78	7	152.3	149.4	0.79	0.57	0.31	0.14	0.08
	89	8	171.8	170.0	0.77	0.55	0.29	0.13	0.06
	100	9	194.7	190.5	0.78	0.58	0.34	0.20	0.12
1023 (10)	37	3	75.5	77.16	0.72	0.46	0.21	0.06	0.02
	49	4	100.9	102.1	0.71	0.49	0.23	0.09	0.04
	73	6	152.7	151.7	0.78	0.53	0.26	0.11	0.06
	85	7	178.5	176.4	0.77	0.55	0.28	0.11	0.07
	97	8	204.8	201.0	0.82	0.59	0.31	0.12	0.07
	109	9	230.0	225.6	0.80	0.60	0.32	0.14	0.08
Mean Estimated α-levels					0.74	0.50	0.25	0.11	0.06

28.69 (and provides better correspondence to the simulated distribution of likelihood ratios). The other example arises with 6 inputs and 7 hidden units, where the estimated value of 58.92 would be replaced by 61.11 (which is slightly further away from the simulated result of 59.6).

Simulation Results for Continuous Input Variables

Results shown in Table 5 were generated using independently distributed continuous inputs from a standard normal distribution. All neural network models were fit using a weight decay of 0.01. Table 5 shows that the simulated distribution of the likelihood ratio test for independence again closely followed a chi-square distribution. In a large percentage of the cases, all of the simulated α-levels were within a few percent of the expected percentiles. No systematic differences were evident in the results. Figures 4a and 4b show two examples where: (1) the simulated distribution varied a few percent from the expected percentiles (2 inputs and 2 hidden units), and (2) the simulated distribution fell extremely close to the corresponding chi-square distribution (2 inputs and 10 hidden units). Both figures show noticeable variability in at least one extreme end of the distribution; however, these few points have little effect on the resulting significance levels that would be of practical interest.

Table 6 shows additional results for other cases with 3 or 8 hidden units and up to 10 inputs. Results again showed simulated α-levels very close to the expected percentiles. Q-Q plots (not shown here) showed similar findings as in Figures 4a and 4b.

Estimating Degrees of Freedom for Continuous Input Variables

As opposed to the case of binary inputs, the degrees of freedom (*df*) for continuous input variables do not have any specific limiting value. Therefore, it was assumed that the *df* would be a continuous (and probably linear) function of the number of hidden units. Further, it was assumed that the result would increase by some constant amount with an increase in the number of input variables. Using the results in Table 5, the relationship $df = h \times [3 \times (k-2) + 5]$ is obtained, which appears to hold well across those results (with 2, 5, and 10 hidden units). Since the specific values from Table 5 were not used to derive this relationship (other than observing the general trend), subsequent results combine simulations from Tables 5 and 6 (i.e., 2-10 inputs and 2, 3, 5, 8 and 10 hidden units). Figure 5 shows the relationship between the simulated and estimated *df* from the results in Tables 5 and 6. The plot illustrates a close correspondence between the simulated and estimated results, especially for smaller degrees of freedom.

Results in Table 7 show somewhat greater variability in the *df* and subsequent significance levels. Only the 5% significance level showed no systematic error, with most of the simulations giving a result (for 5% significance) within 2% of the correct level (e.g., between 3% and 7%). The variability in significance levels can be attributed to either the difference between the simulated and estimated *df* and/or the variability from a chi-square distribution. In most cases, the estimated *df* was at least slightly higher than the simulated result.

Simulation Results for both Binary and Continuous Input Variables

Table 8 shows results for both binary and continuous input variables. For each of these simulations, the number of hidden units was kept constant at 2, the number of continuous inputs was specified as 2, 5, or 10, and the number of binary inputs was varied between 2 and 10. The degrees of freedom (*df*) in parentheses in the first two columns of the table are the estimated values for model complexity (as described in the previous sections of this report). The additional *df* (column 5) gives the difference between the simulated *df* (when combining a given number of continuous and binary inputs) and the sum of estimated *df* (totaled from columns 1, 2 and 3).

The results in Table 8 illustrate several key issues. First, the simulation results show substantially more variability than predicted by the chi-square distribution, which is most likely a consequence of sub-optimal results from the minimization (of the deviance) routine in S-Plus. Secondly, a definite trend exists between the

number of continuous and binary variables and the additional *df* gained (or lost) when combining a given number of (continuous and binary) input variables.

At this point, the observed trends could be used to derive estimates of model complexity for the cases in Table 8 and for other cases with larger numbers of hidden units and other combinations of continuous and binary inputs (as done previously when separately considering continuous or binary inputs). However, the lack of correspondence with the chi-square distribution, and the subsequent need for improved model fitting (e.g., more global optimization procedures) would invalidate any subsequent findings. Therefore, modifications of the S-Plus procedures need to be pursued for these cases before any specific rules can be effectively formulated for the case of both continuous and binary inputs.

Table 5: Likelihood Ratio Statistic for Model Independence with Continuous Inputs

Inputs	Hidden Units	# Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	2	9	10.8	25.0	0.74	0.51	0.25	0.08	0.03
	5	21	26.2	60.5	0.75	0.51	0.23	0.08	0.04
	10	41	49.8	99.4	0.75	0.50	0.25	0.09	0.05
3	2	11	16.6	33.6	0.74	0.51	0.26	0.10	0.04
	5	26	41.0	71.2	0.74	0.52	0.26	0.11	0.06
	10	51	82.5	171.2	0.76	0.50	0.24	0.10	0.05
4	2	13	22.2	40.0	0.77	0.53	0.25	0.09	0.04
	5	31	55.4	111.0	0.75	0.51	0.27	0.10	0.04
	10	61	115.3	216.9	0.75	0.49	0.26	0.11	0.06
5	2	15	27.7	57.9	0.77	0.51	0.25	0.07	0.03
	5	36	69.3	131.4	0.75	0.51	0.26	0.09	0.04
	10	71	144.8	293.4	0.75	0.50	0.24	0.10	0.05
6	2	17	33.4	65.4	0.76	0.54	0.27	0.08	0.04
	5	41	83.0	164.3	0.74	0.50	0.25	0.10	0.05
	10	81	176.7	341.1	0.74	0.53	0.24	0.09	0.05
7	2	19	38.7	100.1	0.77	0.51	0.21	0.07	0.02
	5	46	98.3	202.0	0.75	0.50	0.27	0.08	0.04
	10	91	205.2	375.8	0.75	0.51	0.25	0.11	0.05
8	2	21	44.8	101.1	0.78	0.52	0.23	0.08	0.04
	5	51	112.5	220.8	0.74	0.49	0.24	0.11	0.06
	10	101	239.0	476.9	0.75	0.51	0.25	0.10	0.04
9	2	23	49.9	142.2	0.79	0.53	0.19	0.05	0.02
	5	56	127.4	239.9	0.74	0.48	0.26	0.11	0.06
	10	111	269.0	487.6	0.73	0.50	0.27	0.11	0.05
10	2	25	54.6	166.1	0.80	0.49	0.19	0.05	0.03
	5	61	140.8	280.2	0.76	0.49	0.24	0.12	0.06
	10	121	299.5	546.4	0.75	0.51	0.26	0.10	0.04
Mean Simulated α-levels					0.75	0.51	0.25	0.09	0.04

Table 6: Additional Simulated Likelihood Ratio Statistics with Continuous Inputs

Inputs	Hidden Units	# Parameters	Likelihood Ratio		Simulated α-levels				
			Mean	Variance	0.75	0.50	0.25	0.10	0.05
2	3	13	15.9	35.4	0.76	0.51	0.23	0.08	0.04
	8	33	40.3	82.8	0.76	0.51	0.23	0.10	0.04
3	3	16	24.4	42.0	0.74	0.51	0.29	0.10	0.05
	8	41	65.9	135.7	0.73	0.49	0.25	0.11	0.06
4	3	19	32.9	60.6	0.73	0.53	0.27	0.10	0.05
	8	49	90.5	171.1	0.75	0.49	0.25	0.10	0.06
5	3	22	41.2	85.7	0.76	0.53	0.26	0.07	0.04
	8	57	115.2	200.1	0.73	0.51	0.26	0.11	0.06
6	3	25	48.9	84.4	0.73	0.51	0.29	0.13	0.05
	8	65	139.3	231.3	0.72	0.49	0.29	0.13	0.06
7	3	28	57.0	100.7	0.75	0.52	0.24	0.12	0.06
	8	73	160.9	299.5	0.73	0.49	0.27	0.11	0.06
8	3	31	64.9	140.9	0.75	0.50	0.25	0.10	0.04
	8	81	187.3	376.1	0.75	0.50	0.24	0.10	0.05
9	3	34	74.3	158.7	0.78	0.47	0.24	0.09	0.05
	8	89	211.4	421.9	0.75	0.50	0.25	0.11	0.05
10	3	37	81.4	171.6	0.76	0.50	0.22	0.09	0.06
	8	97	235.5	392.1	0.74	0.50	0.27	0.13	0.06
Mean Simulated α-levels					0.75	0.50	0.25	0.10	0.05

Figure 5: Plot of the Estimated by Simulated Degrees Of Freedom for Continuous Inputs and 2, 5 and 10 Hidden Units

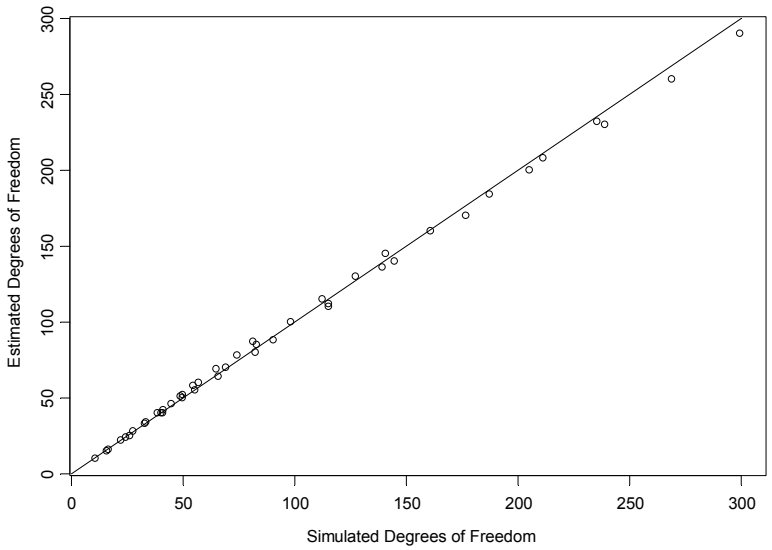


Table 7: Estimated and Simulated Degrees of Freedom with Continuous Inputs

Inputs	Hidden Units	# Parameters	Simulated *df*	Estimated *df*	Simulated α-levels using the Estimated *df*				
					0.75	0.50	0.25	0.10	0.05
2	2	9	10.8	10	0.79	0.58	0.32	0.11	0.05
	3	13	15.9	15	0.80	0.57	0.28	0.11	0.05
	5	21	26.2	25	0.80	0.58	0.28	0.11	0.06
	8	33	40.3	40	0.77	0.53	0.24	0.11	0.04
	10	41	49.8	50	0.75	0.49	0.25	0.08	0.05
3	2	11	16.6	16	0.77	0.55	0.30	0.13	0.06
	3	16	24.4	24	0.76	0.54	0.31	0.12	0.06
	5	26	41.0	40	0.77	0.57	0.30	0.13	0.07
	8	41	65.9	64	0.78	0.55	0.31	0.15	0.08
	10	51	82.5	80	0.81	0.58	0.30	0.14	0.07
4	2	13	22.2	22	0.77	0.54	0.26	0.10	0.04
	3	19	32.9	33	0.73	0.53	0.27	0.10	0.05
	5	31	55.4	55	0.76	0.52	0.28	0.10	0.05
	8	49	90.5	88	0.81	0.56	0.31	0.13	0.09
	10	61	115.3	110	0.85	0.62	0.39	0.20	0.11
5	2	15	27.7	28	0.76	0.49	0.24	0.07	0.03
	3	22	41.2	42	0.73	0.49	0.23	0.07	0.03
	5	36	69.3	70	0.73	0.49	0.25	0.09	0.04
	8	57	115.2	112	0.79	0.59	0.34	0.16	0.09
	10	71	144.8	140	0.83	0.62	0.34	0.16	0.09
6	2	17	33.4	34	0.73	0.50	0.24	0.07	0.03
	3	25	48.9	51	0.65	0.43	0.22	0.09	0.03
	5	41	83.0	85	0.68	0.44	0.20	0.07	0.03
	8	65	139.3	136	0.79	0.60	0.36	0.18	0.08
	10	81	176.7	170	0.84	0.67	0.36	0.17	0.10
7	2	19	38.7	40	0.72	0.45	0.17	0.05	0.01
	3	28	57.0	60	0.65	0.40	0.16	0.07	0.04
	5	46	98.3	100	0.71	0.46	0.23	0.06	0.03
	8	73	160.9	160	0.75	0.50	0.29	0.13	0.06
	10	91	205.2	200	0.83	0.61	0.34	0.17	0.08
8	2	21	44.8	46	0.74	0.47	0.19	0.06	0.03
	3	31	64.9	69	0.63	0.36	0.15	0.05	0.01
	5	51	112.5	115	0.69	0.43	0.19	0.08	0.04
	8	81	187.3	184	0.80	0.57	0.30	0.14	0.07
	10	101	239.0	230	0.86	0.67	0.39	0.19	0.09
9	2	23	49.9	52	0.73	0.45	0.14	0.03	0.01
	3	34	74.3	78	0.68	0.36	0.16	0.05	0.02
	5	56	127.4	130	0.69	0.41	0.21	0.09	0.04
	8	89	211.4	208	0.80	0.57	0.31	0.14	0.07
	10	111	269.0	260	0.84	0.65	0.41	0.20	0.11
10	2	25	54.6	58	0.70	0.37	0.11	0.03	0.01
	3	37	81.4	87	0.61	0.34	0.12	0.04	0.02
	5	61	140.8	145	0.68	0.40	0.17	0.07	0.04
	8	97	235.5	232	0.79	0.57	0.32	0.17	0.09
	10	121	299.5	290	0.86	0.66	0.40	0.19	0.10
Mean Simulated α-levels					0.76	0.52	0.27	0.11	0.05

Table 8: Likelihood Ratios with Continuous and Binary Inputs and 2 Hidden Units

Continuous Inputs (*df*)	Binary Inputs	*df*	Mean Likelihood Ratio	Additional *df*	Simulated α-levels				
					0.75	0.50	0.25	0.10	0.05
2 (10)	2	3.0	19.7	6.7	0.73	0.52	0.27	0.10	0.05
	3	7.0	24.8	7.8	0.76	0.51	0.26	0.08	0.04
	4	10.5	31.0	10.5	0.75	0.54	0.24	0.08	0.03
	5	18.5	36.7	8.2	0.78	0.53	0.25	0.06	0.03
	6	24.7	42.4	7.7	0.78	0.54	0.22	0.06	0.02
	7	31.0	47.7	6.7	0.77	0.51	0.22	0.06	0.02
	8	37.6	54.0	6.8	0.78	0.49	0.22	0.07	0.02
	9	44.6	58.4	3.8	0.81	0.49	0.19	0.06	0.02
	10	52.2	64.1	1.9	0.80	0.47	0.18	0.04	0.02
5 (28)	2	3.0	38.0	7.0	0.76	0.51	0.23	0.06	0.02
	3	7.0	43.6	8.6	0.79	0.53	0.20	0.05	0.02
	4	10.5	47.9	9.4	0.80	0.47	0.20	0.05	0.02
	5	18.5	52.9	6.4	0.80	0.50	0.18	0.06	0.02
	6	24.7	59.5	6.8	0.82	0.48	0.17	0.06	0.02
	7	31.0	64.3	5.3	0.84	0.44	0.18	0.04	0.02
	8	37.6	69.5	3.9	0.79	0.47	0.19	0.05	0.01
	9	44.6	73.6	1.0	0.80	0.46	0.19	0.05	0.02
	10	52.2	79.1	-1.1	0.82	0.45	0.17	0.04	0.01
10 (58)	2	3.0	64.4	3.4	0.81	0.46	0.18	0.05	0.02
	3	7.0	69.4	4.4	0.82	0.47	0.18	0.04	0.01
	4	10.5	72.2	3.7	0.80	0.50	0.18	0.05	0.02
	5	18.5	78.1	1.5	0.78	0.47	0.19	0.06	0.03
	6	24.7	83.0	0.3	0.79	0.46	0.16	0.06	0.02
	7	31.0	89.1	0.1	0.78	0.48	0.17	0.04	0.02
	8	37.6	94.7	-1.1	0.78	0.46	0.17	0.05	0.01
	9	44.6	98.6	-4.0	0.79	0.46	0.20	0.05	0.01
	10	52.2	103.3	-6.9	0.79	0.45	0.15	0.04	0.01
Mean Simulated α-levels					0.79	0.49	0.20	0.06	0.02

Categorical Input Variables

Additional simulations were conducted for categorical variables. In theory, a categorical variable with 3 levels should produce fewer degrees of freedom than 2 binary inputs, since the 3 levels would be coded as 2 binary inputs, but would not have an interaction between the 2 levels. Simulations (not shown here) provided evidence of this type of relationship, but simulation results differed substantially from the expected chi-square distribution. Therefore, as in the cases of both binary and continuous inputs, further work on these types of data are being delayed until a better optimization routine can be implemented with S-Plus or with another programming language.

Conclusions

One issue that was not addressed was the correlation of the input variables. All simulations were run with independently generated data. Comparing the current findings to previous analyses with some overlap (Landsittel, et al., 2003) indicates that the

degrees of freedom (*df*) may be somewhat lower with moderately correlated data, which is somewhat intuitive since correlated variables, to some degree, add less information than independently distributed variables. Rather than consider this further complication, it was decided that all simulations should use independent data as a starting point, and the effects of correlation should be addressed separately in a future manuscript.

Another limitation encountered here was the failure of the S-Plus routine to achieve acceptably optimal results in minimizing the deviance. Because the simulations with only binary, or only continuous inputs led to close correspondence with the chi-square distribution (which allows the use of the mean as the *df*), it would be expected that this would hold for models with both binary and continuous inputs. The failure to achieve this result is most likely a function of the (only locally optimal) routines. Future work will address this point through investigating other optimization routines (e.g., genetic algorithms), and incorporating those routines into the current approaches and methodology.

To the best of our knowledge, these studies are the first to use *df* under the null as a measure of model complexity. Unlike generalized linear models or other standard regression methods, the model complexity may vary substantially for different data sets. In terms of the general applicability of this approach, the complexity under the null may provide a more appropriate penalty for subsequent use in model selection in many scenarios, as higher complexity may be desirable if the true underlying association is highly non-linear. In contrast to a measure such as the generalized *df*, where the complexity tends to increase substantially when fit to data with some observed association, the complexity under the null only penalizing the model for incorrectly fitting non-linearity when none exists. Using an AIC-type statistic with generalized or effective *df*, for example, would highly penalize the neural network model for accurately fitting a highly non-linear association, and likely make it very difficult to select an adequately complex model.

Despite these limitations, the results contribute significantly to our understanding of neural network model complexity by providing explicit equations to quantify complexity under a range of scenarios. Once improved methods are implemented to better optimize more complex models (where there was significant variability from the expected chi-square distribution), the derived equations for *df* can be tested across a much wider range of models. Assuming results hold for other scenarios (to be tested after achieving more global optimization), the estimated *df* can be implemented in practice for model selection via AIC or BIC statistics. Such approaches would serve as a favorable alternative to any of the ad-hoc approaches currently being utilized in practice.

Acknowledgements

The author would like to acknowledge Mr. Dustin Ferris, who, as an undergraduate student in Mathematics at Duquesne University, conducted many of the initial simulations used to generate some of the results in this manuscript, and Duquesne University, for the Faculty Development Fund award, that included funding of Mr. Ferris's stipend.

References

Amari, S.. & Murata, N. (1993). Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, *5*, 140-153.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. NY: Chapman and Hall.

Hodges, J. S., & Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, *88*(*2*), 367-379.

Landsittel, D., Singh, H., & Arena, V. C. (2003). Null distribution of the likelihood ratio statistic with feed-forward neural networks. *Journal of Modern and Applied Statistical Methods*, *1*(*2*), 333-342.

Liu, Y. (1995). Unbiased estimate of generalization error and model selection in neural networks. *Neural Networks*, *8*(*2*), 215-219.

Moody, J. E. (1992). The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems 4*, 847-854. San Mateo, CA: Morgan Kaufmann.

Murata, N., Yoshizawa, S., & Amari, S. (1991). A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen, K Makisara, O. Simula, & J. Kangas (Eds.), *Artificial neural networks*, 9-14. North Holland: Elsevier Science Publishers.

Paige, R. L., & Butler, R. W. (2001). Bayesian inference in neural networks. *Biometrika*, *88*(*3*), 623-641.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, MA: Cambridge University Press.

Tetko, I. V., Villa, A. E., & Livingstone, D. J. (1996). Neural network studies 2: Variable selection. *Journal of Chemical Informatics and Computer Science*, *36*(*4*), 794-803.

Venables, W. N., & Ripley, B. D. (1997). *Modern applied statistics with S-Plus*. NY: Springer-Verlag.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, *93*(*441*), 120-131.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 505-510

1538 – 9472/09/$95.00

Multiple Search Paths and the General-To-Specific Methodology

Paul Turner
Loughborough University
United Kingdom

Increased interest in computer automation of the general-to-specific methodology has resulted from research by Hoover and Perez (1999) and Krolzig and Hendry (2001). This article presents simulation results for a multiple search path algorithm that has better properties than those generated by a single search path. The most noticeable improvements occur when the data contain unit roots.

Key words: Multiple search paths, general-to-specific, Monte Carlo simulation.

Introduction

The general-to-specific methodology introduced by Davidson, et al. (1978) and discussed by Gilbert (1986) is now a well established part of the technical toolkit of applied econometricians. The idea of this approach is to begin with a deliberately over-parameterized model, examine its properties (particularly those of the residuals) to ensure that it is data congruent and then to progressively simplify the model to obtain a parsimonious specification. Arguably, the main advantage of this approach is that, provided the original over-parameterized model is data congruent, tests of restrictions are always conducted against a statistically well specified alternative model. This contrasts with the alternative specific-to-general approach in which the alternative model is frequently badly specified, thereby invalidating the testing procedure.

A typical situation facing a modeler can be illustrated as follows. The modeler begins with a general model of the form which relates two variables of interest y and x which follow a dynamic relationship disturbed by a random error u:

$$y_t = \beta_0 + \sum_{i=0}^{4} \beta_{i+1} x_{t-i} + \sum_{i=1}^{4} \beta_{5+i} y_{t-i} + u_t \quad (1)$$

Economic theory suggests that an equilibrium relationship exists between the variables y and x. However, theory typically indicates little about the short-run dynamic relationship between the variables. Therefore, beginning with (1)[1], exclusion restrictions for the right-hand side variables are tested and these are progressively eliminated until either the null $H_0 : \beta_i = 0$ is rejected or the model begins to show signs of misspecification in the form of serial correlation in the residuals, heteroscedasticity, non-normality etc. When a parsimonious specification is found then the model is often re-written in a more convenient form such as the error-correction representation.

One of the problems which arises with the general-to-specific methodology is the search path involved in moving from the general model (1) to a parsimonious specification is not unique. Typically, the general model contains a large number of highly co-linear variables. Exclusion of large numbers of variables at an early stage is a dangerous strategy since variables that may be insignificant in the most general model may become significant as other co-linear variables are excluded. Most advocates of this methodology therefore recommend proceeding gradually, eliminating a few variables at each stage of the specification search, until the final specification is obtained. However, the number of possible search paths

Paul Turner is Reader in Economics in the School of Business and Economics, Loughborough University. Email him at: P.M.Turner@lboro.ac.uk

may become large even in a relatively small model. Suppose, for example, that the true model requires a total of n restrictions on the most general model. It follows that there are $n!$ separate search paths which involve the elimination of one variable at each stage and which will succeed in getting from the general model to the final, correct specification. If the elimination of several variables at once is allowed then the number of search paths increases still further.

Another problem arising within the general-to-specific methodology is that there is always the chance of making a Type II error during the process of the specification search and a variable which should be present in the final specification is eliminated at some stage. The result of this is typically that other variables, which ideally would have been excluded in the final specification, are retained as proxies for the missing variable. The resulting final specification is therefore over-parameterized. It is difficult to identify cases such as this from real world data where the investigator does not have the luxury of knowledge of the data generation process. However, it is straightforward to demonstrate this phenomenon using Monte Carlo analysis of artificial data sets.

In the early years of general-to-specific analysis it was argued that the only solution to the problems discussed above was to rely on the skill and knowledge of the investigator. For example, Gilbert (1986) argued the following:

> How should the econometrician set about discovering congruent simplifications of the general representation of the DGP …. Scientific discovery is necessarily an innovative and imaginative process, and cannot be automated. (p.295)

However, more recent research by Hoover and Perez (1999), Hendry and Krolzig (2001) and Krolzig and Hendry (2001) has suggested that automatic computer search algorithms can be effective in detecting a well specified econometric model using the now established 'general-to-specific' methodology. This has been facilitated by the introduction of the PC-GETS computer package which will automatically conduct a specification search to obtain the best data congruent model based on a given data set.

The purpose of this paper is to investigate the properties of a simple automatic search algorithm in uncovering a correctly specified parsimonious model from an initially overparameterized model. The algorithm works by estimating multiple search paths and choosing the final specification which minimizes the Schwartz criterion. This is compared with a naïve search algorithm in which the least significant variable in the regression is successively eliminated until all remaining variables are significant at a pre-determined level.

Methodology

The main problem encountered in conducting multiple search paths is the number of possible search paths that might be legitimately investigated. For example, consider a model in which the final specification involves twelve exclusion restrictions relative to the original model (not an unusual situation when working with quarterly data). In this case there are $12! = 479,001,600$ possible search paths involving the progressive elimination of one variable at each stage. Therefore, even with the power of modern computing, consideration of every possible search path is simply not an option. However, the situation is not as impossible as it may first appear. Many search paths will eventually converge on the same final specification and the problem is simply to ensure that enough are tried so as to maximize the chance of obtaining the correct specification. The pseudo-code below sets out the algorithm used in this research to achieve this.

FOR j = 1 to R, where R is a predetermined number of iterations.

REPEAT UNTIL $\left|t_{\hat{\beta}_i}\right| > t_c$ where t_c is a predetermined critical value for all $i = 1,..,N$ where N is the number of variables included in the equation.

Estimate equation.

FOR each variable in the model examine $\left|t_{\hat{\beta}_i}\right|$. IF $\left|t_{\hat{\beta}_i}\right| < t_c$ AND $\gamma > 0.5$ where γ is a random drawing from a uniform distribution with the interval $[0,1]$ THEN eliminate associated variable and re-estimate equation. ELSE IF $\gamma < 0.5$ THEN retain variable.

IF $\left|t_{\hat{\beta}_i}\right| > t_c$ for all *i* then STOP and record the variables included in the equation as well as the value of the Schwartz criterion. Otherwise go back to previous step.

FOR j = 1 to R, compare the value of the Schwartz criterion for each final specification and choose the specification with the lowest value

The data generation process takes the form of the familiar partial adjustment model. This formulation is consistent with a cost minimization process in which agents minimize a quadratic cost function which includes costs of adjustment as well as costs of being away from equilibrium. The equation used to generate the data takes the form:

$$y_t = 0.5x_t + 0.25y_{t-1} + u_t \qquad t:1,...,T \tag{2}$$

where $u_t : t = 1,...,T$ are iid standard normal random variables. The *x* variable is generated in two alternative ways. In the first $x_t : t = 1,...,T$ are also iid standard normal random variables with $\text{cov}(x_t, u_t) = 0$. In the second, $x_t = x_{t-1} + \varepsilon_t$ where $\varepsilon_t : t = 1,...,T$ are iid standard normal variables with $\text{cov}(x_t, \varepsilon_t) = 0$. Thus in case 1 the relationship is one between stationary variables while, in case 2, it is between $I(1)$ variables.

Using (2) to generate the data and (1) as the starting point for a specification search, the search algorithm discussed above is applied as well as the naïve search algorithm of simply eliminating the least significant variable at each stage of the search process. Ten thousand specification searches[2] are carried out using seeded pseudo-random numbers generated by the EViews regression package and the results of each search are classified according to the classification set out by Hoover and Perez (1999) as shown below:

A: Final model = True Model
B: True Model $\subset$ Final Model and $\hat{\sigma}_{Final} < \hat{\sigma}_{True}$
C: True Model $\subset$ Final Model and $\hat{\sigma}_{Final} > \hat{\sigma}_{True}$
D: True Model $\not\subset$ Final Model and $\hat{\sigma}_{Final} < \hat{\sigma}_{True}$
E: True Model $\not\subset$ Final Model and $\hat{\sigma}_{Final} > \hat{\sigma}_{True}$

Thus the final specification is classified as to whether it matches the true model (case A), contains all the variables included in the true model and has a lower standard error (case B), contains all the variables included in the true model but has a higher standard error (case C), omits at least one variable from the true model but has a lower standard error (case D) or omits at least one variable from the true model and has a higher standard error (case E).

Results

Table 1 presents the results for the multiple search path algorithm when the data are stationary. In all cases $R = 100$, that is 100 different specification searches were carried out and the equation with the lowest Schwartz criterion[3] was chosen. Examination of Table 1 indicates that both the sample size and the choice of critical value used in the specification search are important factors. If the sample size is small $T = 100$ then $t_c = t_c^{5\%}$ performs better than $t_c = t_c^{1\%}$ value in terms of identifying the true model more often (case A) and avoiding the elimination of variables that should be present in the true model (case E). However, as the sample size increases, this situation is reversed and in large samples with $T = 500$ then $t_c = t_c^{1\%}$ performs much better than $t_c = t_c^{5\%}$. Note that case C is never observed in any of the simulations carried out.

Does the multiple search path algorithm offer any gains over a naïve specification search? Examination of the results in Table 2 suggests that this is the case. In all cases the multiple search path algorithm identifies the true

model more often. Moreover, as the sample size gets large, the frequency with which the multiple search path algorithm identifies the true model appears to be converging towards 100% with $t_c = t_c^{1\%}$. This is not the case for the naïve algorithm in which, with the same specification, the true model was identified in only 67.6% of the simulations.

Next, the effects of working with non-stationary data was considered. Here the *x* variable is generated as a random walk series with the implication that the *y* variable also contains a unit root. However, the specification of an equilibrium relationship between the variables ensures that they are co-integrated. This means that it is still reasonable to conduct a specification search in levels of the series even though individually each series contains a unit root. The results for the multiple search path algorithm are given in Table 3.

The results from Table 3 are very similar to those for non-stationary data shown in Table 1. The actual percentages differ slightly but the general pattern remains the same. If the sample size is small then $t_c = t_c^{5\%}$ performs better than $t_c = t_c^{1\%}$. However, as the sample size gets larger, this situation is reversed with case A converging towards 100% (when $t_c = t_c^{1\%}$) as the sample size becomes large.

Table 1: Multiple Search Paths General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, *x* and *u* are independently generated *iid* processes)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	52.4	48.2	76.4	83.3	80.9	93.0
B	15.2	4.0	17.3	5.7	19.1	6.9
C	0.0	0.0	0.0	0.0	0.0	0.0
D	14.7	10.8	2.8	2.5	0.0	0.0
E	17.7	37.0	3.5	8.5	0.0	0.0

Table 2: Single Search Path General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, *x* and *u* are independently generated *iid* processes)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	36.5	34.2	49.4	60.3	51.4	67.6
B	18.5	3.3	27.5	6.0	29.4	7.1
C	0.0	0.0	0.0	0.0	0.0	0.0
D	24.3	17.4	12.2	14.9	9.6	12.9
E	20.7	45.1	10.9	18.8	9.6	12.4

Finally, the multiple search path algorithm is contrasted with the naïve algorithm for the case of non-stationary data. The results for the naïve algorithm are shown in Table 4. These indicate that the naïve algorithm performs extremely badly when applied to non-stationary data. Case A is achieved in at best one quarter of the simulations, even with a large sample $T = 500$ and irrespective of the critical value employed. This suggests that the real value of a multiple search path algorithm may lie in its application to the modeling of non-stationary series. Since this is very often the case with econometric model building, it suggests that the approach may have considerable practical value.

Conclusion

In this article the use of a multiple search path algorithm for the general-to-specific approach to econometric analysis has been investigated. It has been shown that this algorithm has significant advantages over a naïve approach to specification searches. Moreover the relative advantage of this approach increases when dealing with non-stationary data. Since non-stationary data is the norm rather than the exception in econometric model building, it is arguable that a multiple search path approach offers real advantages to the applied econometrician.

Table 3: Multiple Search Paths General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, x is a random walk process and u is a stationary *iid* process)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	54.4	49.8	81.2	85.2	88.9	94.7
B	10.1	2.7	11.5	4.4	11.1	5.2
C	0.0	0.0	0.0	0.0	0.0	0.0
D	19.8	14.7	4.7	4.7	0.0	0.1
E	15.8	32.9	2.6	5.7	0.0	0.0

Table 4: Single Search Path General-To-Specific
($y_t = 0.5x_t + 0.25y_{t-1} + u_t$, x is a random walk process and u is a stationary *iid* process)

	T=100		T=200		T=500	
Classification	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size	5% Nominal Size	1% Nominal Size
A	17.2	16.4	20.9	27.8	14.8	21.5
B	16.7	2.3	29.8	3.9	33.3	7.5
C	0.0	0.0	0.0	0.0	0.0	0.0
D	39.3	32.3	25.5	32.4	26.6	36.9
E	26.8	49.0	23.8	35.9	25.3	34.1

Notes

[1]The lag length in equation (1) is set at 4 for illustrative purposes only. This is often the case when dealing with quarterly data but alternative lag lengths are frequently employed for data with different frequencies.

[2]The specification searches were carried out using an EViews program which is available from the author on request.

[3]In fact, examination of the results indicates that many different specification search paths converge on the true model. The problem is not one of picking a single search path which gives the correct result but rather one of avoiding rogue search paths which give the wrong result.

References

Davidson, J., Hendry, D., Srba, F., & Yeo, S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, *88*, 661-692.

Gilbert, C. L. (1986). Professor Hendry's econometric methodology, *Oxford Bulletin of Economics and Statistics*, *48*, 283-307.

Hendry, D.F., & Krolzig, H. (2001). New developments in automatic general-to-specific modelling. In Stigum, B. (Ed.) *Econometrics and the philosophy of economics*. Cambridge, MA: MIT Press.

Hoover, K., & Perez, S. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, *2*, 1-25.

Krolzig, H., & Hendry, D. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, *25*, 831-866.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 511-519

1538 – 9472/09/$95.00

Closed Form Confidence Intervals for Small Sample Matched Proportions

James F. Reed III
Christiana Care Hospital System,
Newark, Delaware

The behavior of the Wald-z, Wald-c, Quesenberry-Hurst, Wald-m and Agresti-Min methods was investigated for matched proportions confidence intervals. It was concluded that given the widespread use of the repeated-measure design, pretest-posttest design, matched-pairs design, and cross-over design, the textbook Wald-z method should be abandoned in favor of the Agresti-Min alternative.

Key words: Matched proportions, Wald-z, Wald-c, Quesenberry-Hurst, Wald-m, Agresti-Min.

Introduction

Matched-pairs data are common in clinical trials. Study designs that use paired data include the repeated-measure design, pretest-posttest, the matched-pairs design, and the cross-over design. When the response variable is dichotomous and when two dichotomous measurements are available, the data may be summarized as shown in Table 1.

Table 1: Paired Data Study Design Responses

	Test II		
Test I	Success	Failure	Total
Success	a	b	a+b
Failure	c	d	c+d
Total	a+c	b+d	n

For binary responses, McNemar's test is the most commonly applied significance test for comparing the two response distributions. For interval estimation of the difference of proportions, textbooks present the Wald large sample interval (Wald-z). Like the one proportion (Figure 1) and the difference between two independent binomial confidence intervals (Figure 2), the Wald-z interval for matched-pair proportions behaves rather poorly (Figure 3a). Two problems are generally encountered. First, the coverage probability cannot be achieved exactly and secondly, in many instances the Wald-z method does not yield sensible intervals.

The purpose of this study was to investigate the coverage probability of alternative methods for computing confidence intervals to the typical textbook Wald-z or Wald-c (continuity correction). Those alternatives include a simple add four method proposed by Agresti and Min (AM) (2005), a method by Quesenberry and Hurst (QH) (1964), and a modified Wald (Wald-m) suggested by May and Johnson (1998).

Methodology

Notation and Computational Formula

Let $\mathbf{y} = (a, b, c, d)^T$ represent the observed frequencies for a sample from a multinomial distribution with underlying probabilities $\boldsymbol{\pi} = (\pi_a, \pi_b, \pi_c, \pi_d)^T$. Let b be the number of subjects who respond favorably on the first occasion but unfavorably on the second and let c be the number who responds unfavorably on the first occasion but favorably on the second. Let a be the number of subjects who respond favorably on both occasions and let d be the number who respond unfavorably on both occasions; then a + d represents the number of concordant pairs and b + d represents the number of discordant pairs.

James F. Reed III, Ph.D., is a Senior Biostatistician. Email him at: JaReed@ChristianaCare.org.

Figure 1: Coverage Probabilities (n=50) for A Single Proportion Wald Confidence Interval Method

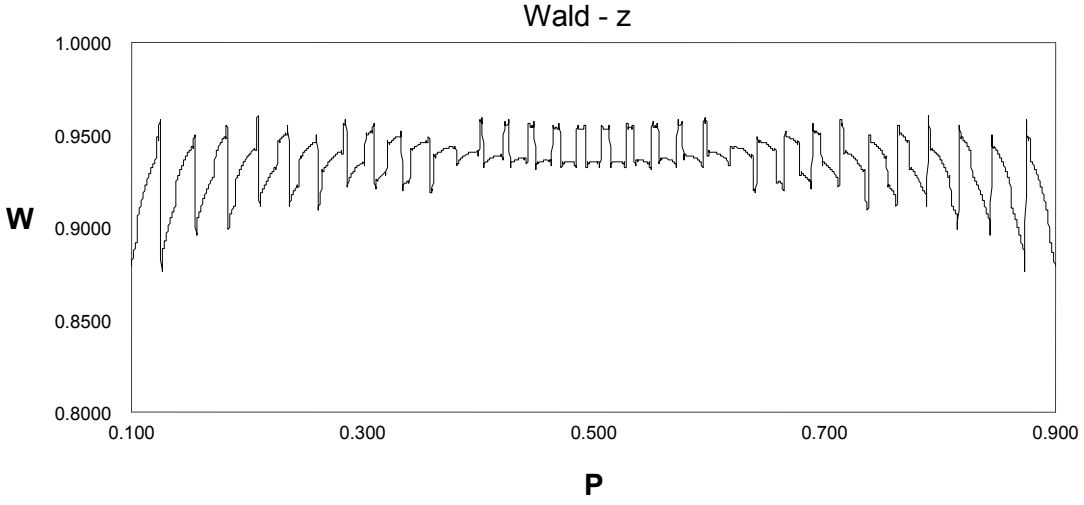
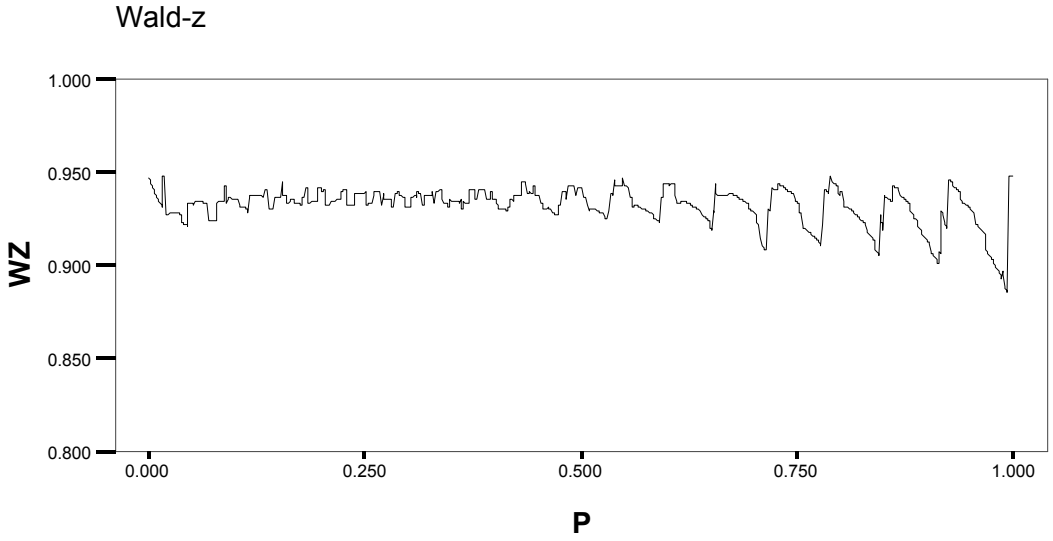


Figure 2: Coverage probabilities for the difference in nominal 95% Wald-z as a function of p_1 when p_2=0.3 with n_1=n_2=20 (Two Independent Proportions)



The confidence interval computational methods based on the data structure given in Table 1 are as follows:

Wald-z:

$$LB = (b-c)/n - z_{\alpha 2}\sqrt{[((b+c)/n-(b-c)^2)/n^2)/n]}$$

and

$$UB = (b-c)/n + z_{\alpha 2}\sqrt{[((b+c)/n-(b-c)^2)/n^2)/n]}.$$

Wald-c:

$$LB = (b-c)/n - \left\{\frac{z_{\alpha 2}\sqrt{[((b+c)/n-(b-c)^2)/n^2)/n]+1}}{n}\right\}$$

and

$$UB = (b-c)/n + \left\{\frac{z_{\alpha 2}\sqrt{[((b+c)/n-(b-c)^2)/n^2)/n]+1}}{n}\right\}.$$

Wald-m:

$$LB = |p_b - p_c| - \{\chi^2\left[(p_b + p_c + 1/n) - (p_b - p_c)^2\right]/n)^{1/2}$$

and

$$UB = |p_b - p_c| + \{\chi^2\left[(p_b + p_c + 1/n) - (p_b - p_c)^2\right]/n)^{1/2}$$

with

$$\chi^2 = \chi^2(\alpha, 1).$$

Quesenberry-Hurst (QH):

$$LB = \frac{(n\,|\,p_b - p_c\,|)}{\left(\frac{(\chi^2 + n) - \left\{\chi^2\left[(\chi^2 + n)(p_b + p_c) - n(p_b - p_c)^2\right]\right\}^{1/2}}{(\chi^2 + n)}\right)}$$

and

$$UB = \frac{(n\,|\,p_b - p_c\,|)}{\left(\frac{(\chi^2 + n) + \left\{\chi^2\left[(\chi^2 + n)(p_b + p_c) - n(p_b - p_c)^2\right]\right\}^{1/2}}{(\chi^2 + n)}\right)}$$

where

$$\chi^2 = \chi^2(\alpha, 1).$$

Agresti-Min (AM):

$$LB = \left(c^* - b^*\right)/n^* - z_{\alpha 2}\sqrt{\left[\left(b^* + c^*\right) - \left(c^* - b^*\right)^2/n^*\right]/n^*}$$

and

$$UB = \left(c^* - b^*\right)/n^* - z_{\alpha 2}\sqrt{\left[\left(b^* + c^*\right) - \left(c^* - b^*\right)^2/n^*\right]/n^*}$$

with

$$b^* = b + 1/2,\ c^* = c + 1/2,\ n^* = n + 2.$$

The joint probability mass function of (Y_b, Y_c) is expressed as a function of Δ [$\Delta = (\pi_c - \pi_b) - (\pi_a - \pi_c) = \pi_b - \pi_c$] and is given by: $f(b,c \mid \Delta, \pi_c) = \Pr(Y_b = b, Y_c = c \mid \Delta, \pi_c) = n!/[b!c!(n-b-c)!]\ (\pi_c + \Delta)^b\ \pi_c^c\ (1 - 2\pi_c - \Delta)^{n-b-c}$. Where, Δ and π_c satisfy the following inequality:

$$\pi_c \in [0, (1-\Delta)/2] \text{ if } 0 < \pi_c < 1,$$

and

$$\pi_c \in [-\Delta, (1-\Delta)/2] \text{ if } -\Delta < \pi_c < 0.$$

Coverage probability (CP) is generally used to evaluate $(1 - \alpha)$ confidence intervals. The coverage probability function CP (Δ) for matched proportions for any Δ is defined as:

$$CP(\Delta) = [\Sigma_k [\Sigma_b \Sigma_c\ I_T(b,c\,|\Delta,\pi_c)\ f(b,c \mid \Delta, \pi_c)]\},$$

where:

$$I_T(b, c \mid \Delta, \pi_c) = 1 \text{ if } \Delta \in [\Delta_l, \Delta_u];\ 0 \text{ otherwise.}$$

Results

The 95% CP (Δ) for $p_c = 0.1$, $n=20$ and $p_b = 0.001, \ldots, 0.999$ for the Wald-z, Wald-c, AM, Wald-m and Quesenberry-Hurst methods are shown in Figure 3.

CP (Δ) probabilities are 0.9125, 0.9545, 0.9401, 0.9435 and 0.0541 respectively. The 95% CP (Δ) for $p_c = 0.25$, $n=30$ and $p_b = 0.001, \ldots, 0.999$ for the Wald-z, Wald-c, AM Wald-m and Quesenberry-Hurst methods are shown in Figure 4.

CP (Δ) probabilities are 0.9334, 0.9611, 0.9425, 0.9484 and 0.9448 respectively. And, the CP (Δ) for $p_c = 0.40$, $n=40$ and $p_b = 0.001, \ldots, 0.999$ for the Wald-z, Wald-c, AM Wald-m and Quesenberry-Hurst methods are shown in Figure 5. CP (Δ) probabilities are 0.9390, 0.9607, 0.9444, 0.9485 and 0.9451 respectively.

The CP (Δ) plots in figures 3-5 demonstrate that the Wald-z method is suboptimal over the range of p, the Wald-c and Wald-m methods are conservative and the Quesenberry-Hurst and Agresti-Min methods are slightly less than nominal.

Conclusion

A number of closed form methods for constructing confidence intervals for paired binary data were proposed. Newcombe (1998) conducted an empirical study to compare the CP (Δ) of ten confidence interval estimators for the difference between binomial proportions based on paired data. He concluded that the profile likelihood estimator and the score test based confidence interval proposed by Tango (1998) performed well in large-sample situations. May

Figure 3: 95% Coverage Probability for Matched Proportions
p_c=0.1, n=20 p_b= 0.001, ..., 1–p_c

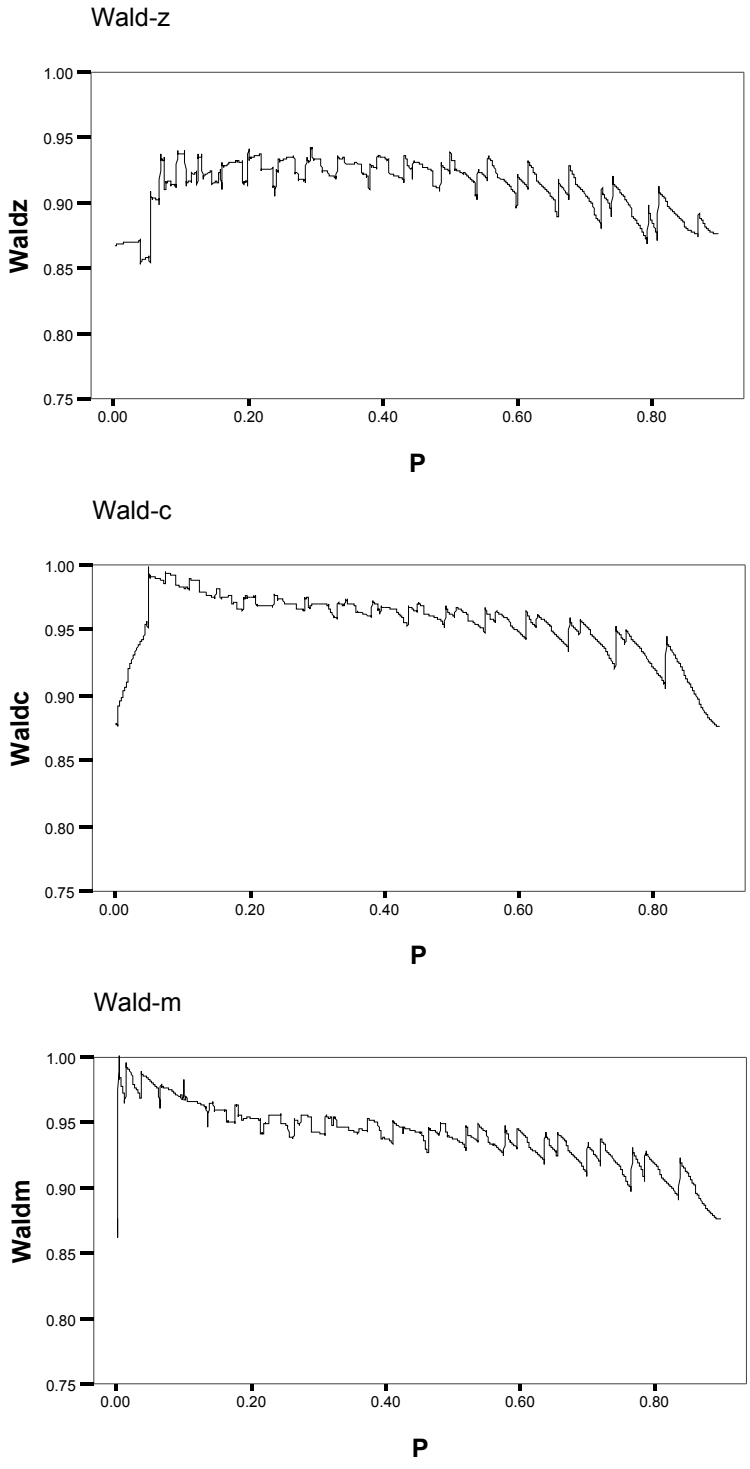


Figure 3 (continued): 95% Coverage Probability for Matched Proportions
p_c=0.1, n=20 p_b= 0.001, ..., 1–p_c

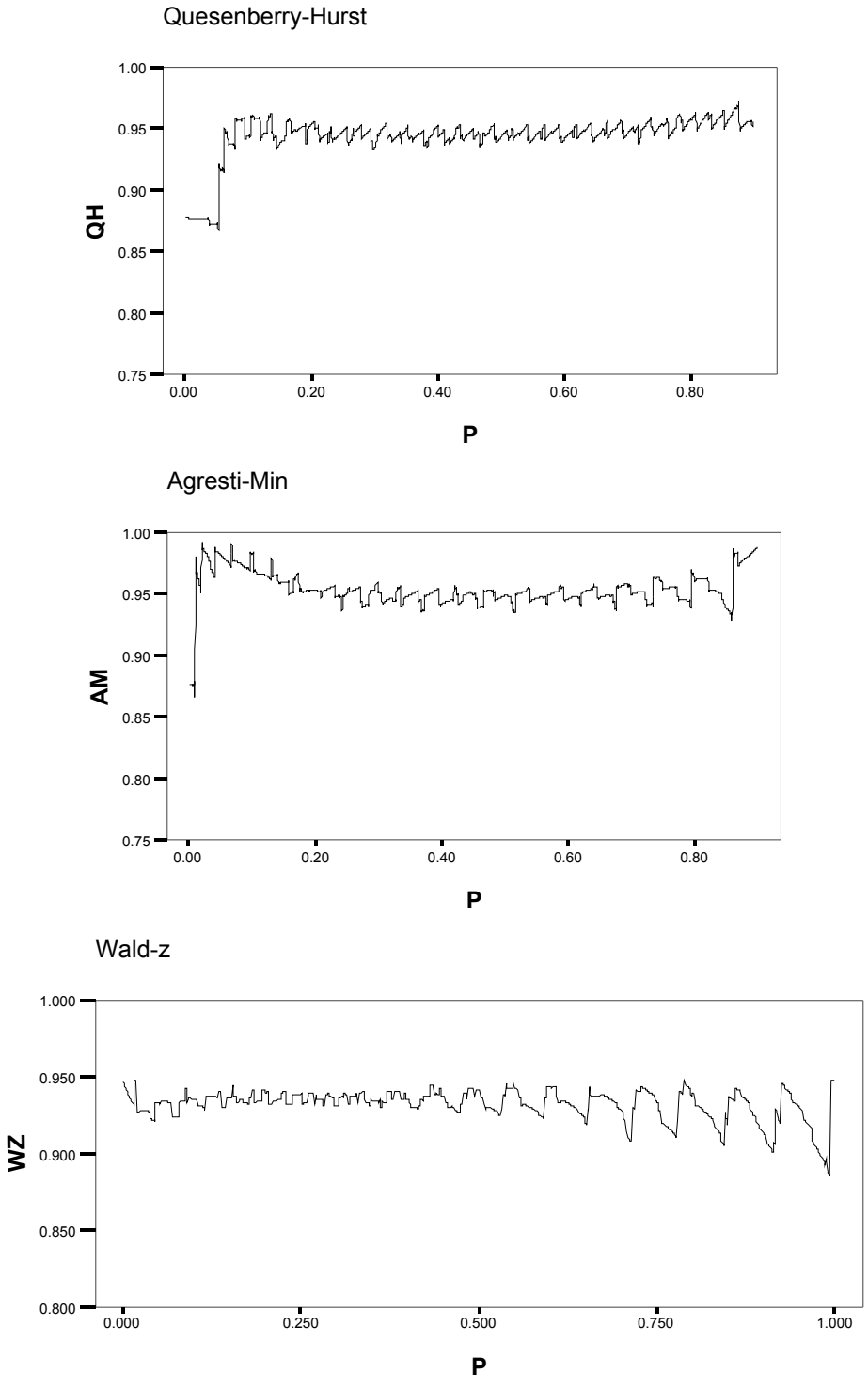


Figure 4: 95% Coverage Probability for Matched Proportions
p_c=0.25, n=30, p_b= 0.001, …, 1–p_c

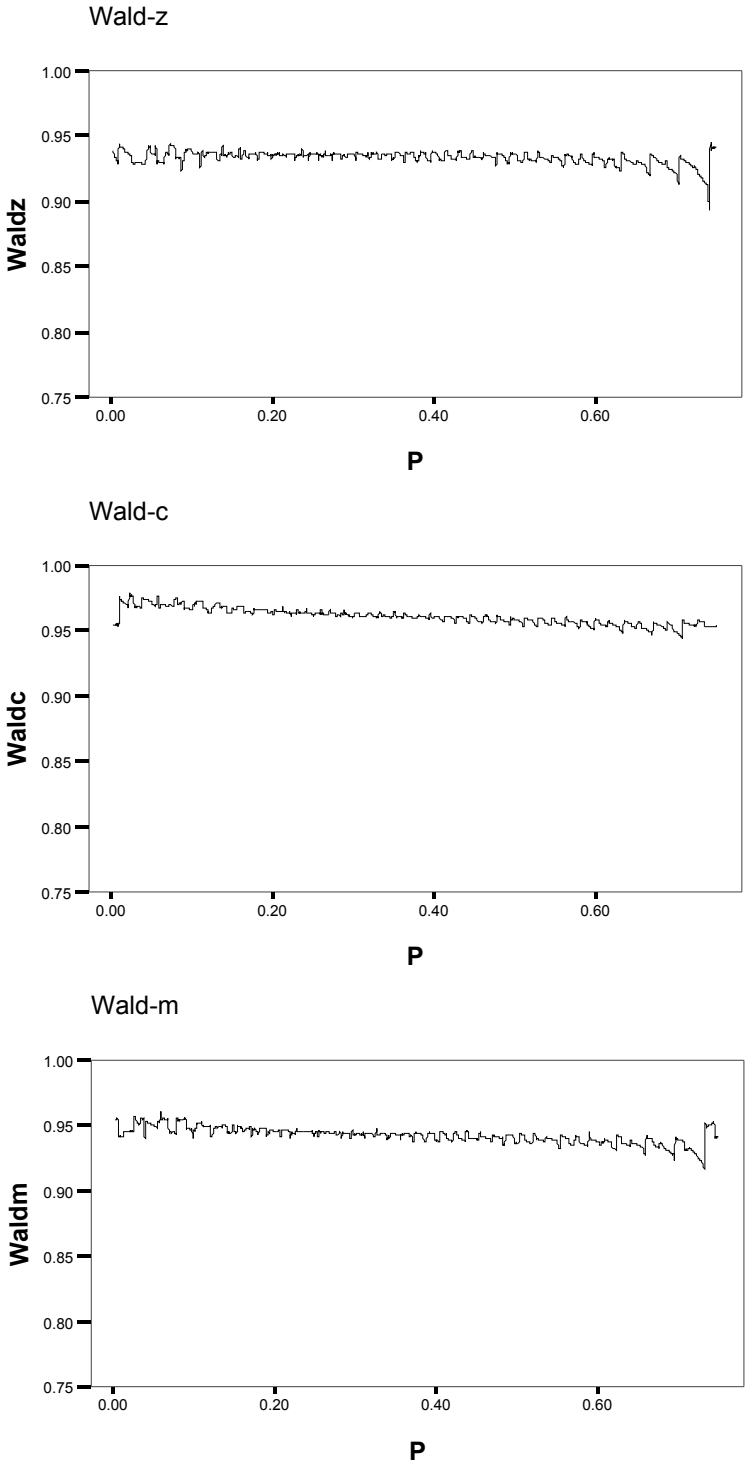


Figure 4 (continued): 95% Coverage Probability for Matched Proportions
p_c=0.25, n=30, p_b= 0.001, …, 1–p_c

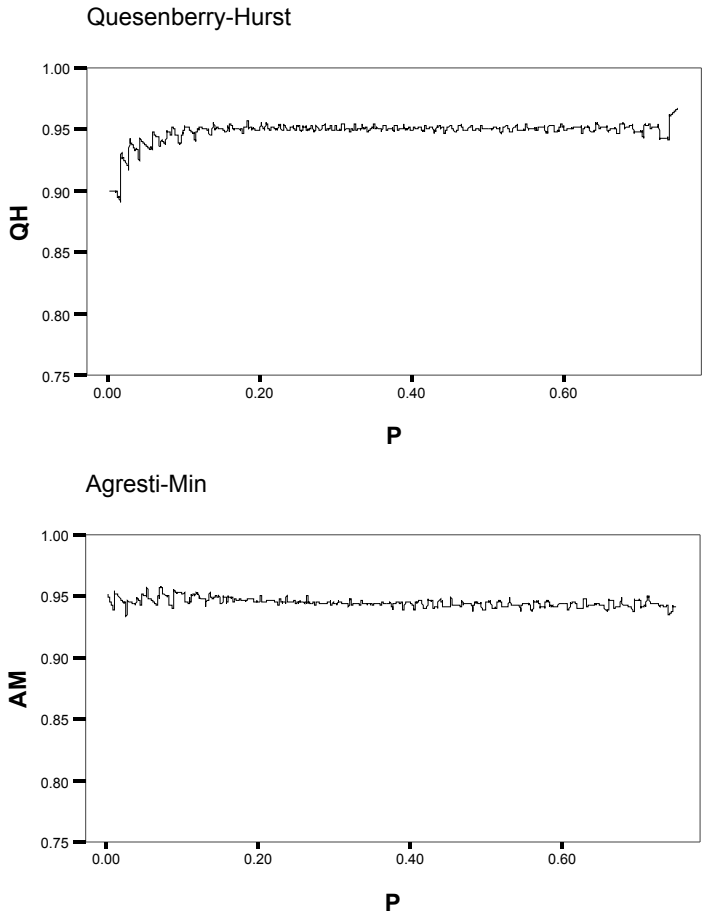


Figure 5: 95% Coverage Probability for Matched Proportions
p_c= 0.4, n=40 and p_b= 0.001, ..., 1 – p_c

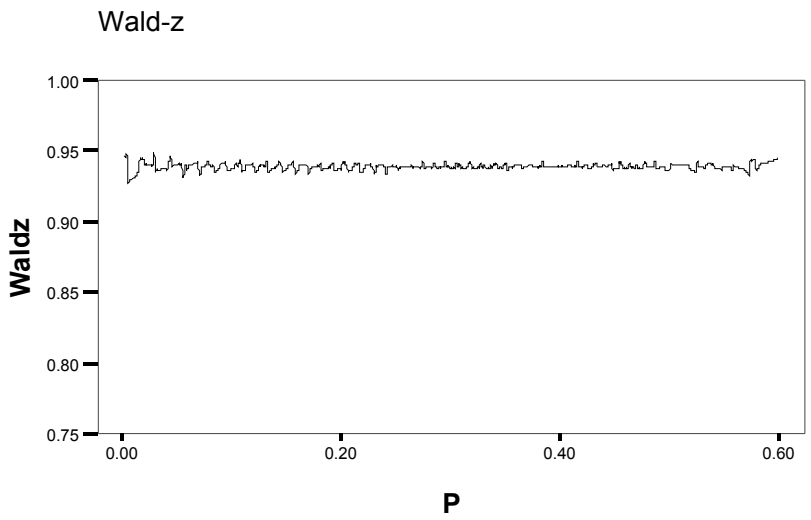


Figure 5 (continued): 95% Coverage Probability for Matched Proportions
p_c= 0.4, n=40 and p_b= 0.001, ..., 1 – p_c

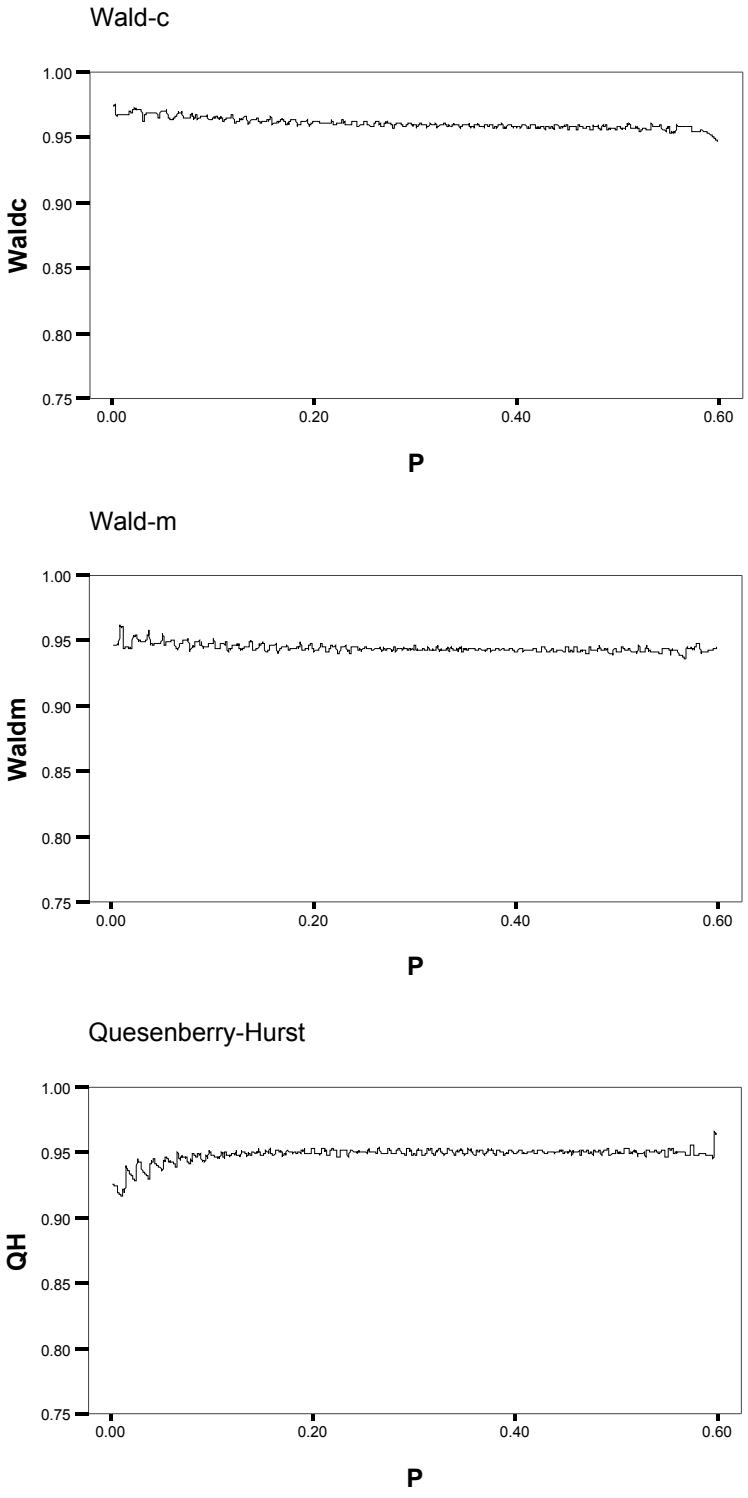
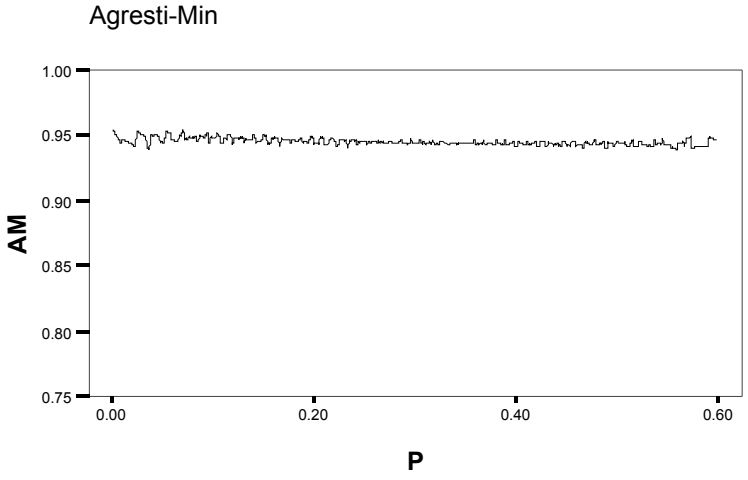


Figure 5 (continued): 95% Coverage Probability for Matched Proportions
p_c= 0.4, n=40 and p_b= 0.001, ..., 1 – p_c



and Johnson (1998), Quesenberry and Hurst (1964) and Agresti and Min (2005) proposed closed form computationally friendly alternatives.

This article focused on constructing confidence intervals using closed form methods for paired data under small-sample designs. In this setting, based on the results, either the Quesenberry-Hurst or Agresti-Min methods are recommended. Given the widespread use of the repeated-measure, pretest-posttest, the matched-pairs, and the cross-over designs, the textbook Wald-z method should be abandoned in favor of either the closed form of Quesenberry-Hurst or Agresti-Min.

References

Agresti, A. & Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*, *24*, 729-740.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, *17*, 2635-2650.

May, W. L., & Johnson, W. D. (1998). Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine*, *16*, 2127-2136.

Quesenberry, C. P., & Hurst, D. C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, *6*, 191-195.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, *17*, 891-908.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 520-525

1538 – 9472/09/$95.00

Confidence Interval Estimation for Intraclass Correlation Coefficient Under Unequal Family Sizes

Madhusudan Bhandary
Columbus State University

Koji Fujiwara
North Dakota State University

Confidence intervals (based on the χ^2-distribution and (Z) standard normal distribution) for the intraclass correlation coefficient under unequal family sizes based on a single multinormal sample have been proposed. It has been found that the confidence interval based on the χ^2-distribution consistently and reliably produces better results in terms of shorter average interval length than the confidence interval based on the standard normal distribution: especially for larger sample sizes for various intraclass correlation coefficient values. The coverage probability of the interval based on the χ^2-distribution is competitive with the coverage probability of the interval based on the standard normal distribution. An example with real data is presented.

Key words: Z-distribution, χ^2-distribution, intraclass correlation coefficient, confidence interval.

Introduction

Suppose, it is required to estimate the correlation coefficient between blood pressures of children on the basis of measurements taken on p children in each of n families. The p measurements on a family provide p(p − 1) pairs of observations (x, y), x being the blood pressure of one child and y that of another. From the n families we generate a total of np(p − 1) pairs from which a typical correlation coefficient is computed. The correlation coefficient thus computed is called an intraclass correlation coefficient. Statistical inference concerning intraclass correlations is important because it provides information regarding blood pressure, cholesterol, etc. in a family within a particular race.

The intraclass correlation coefficient ρ as a wide variety of uses for measuring the degree of intrafamily resemblance with respect to characteristics such as blood pressure, cholesterol, weight, height, stature, lung capacity, etc. Several authors have studied statistical inference concerning ρ based on single multinormal samples (Scheffe, 1959; Rao, 1973; Rosner, et al., 1977, 1979; Donner & Bull, 1983; Srivastava, 1984; Konishi, 1985; Gokhale & SenGupta, 1986; SenGupta, 1988; Velu & Rao, 1990).

Donner and Bull (1983) discussed the likelihood ratio test for testing the equality of two intraclass correlation coefficients based on two independent multinormal samples under equal family sizes. Konishi and Gupta (1987) proposed a modified likelihood ratio test and derived its asymptotic null distribution. They also discussed another test procedure based on a modification of Fisher's Z-transformation following Konishi (1985). Huang and Sinha (1993) considered an optimum invariant test for the equality of intraclass correlation coefficients under equal family sizes for more than two intraclass correlation coefficients based on independent samples from several multinormal distributions.

For unequal family sizes, Young and Bhandary (1998) proposed Likelihood ratio test, large sample Z-test and large sample Z^*-test for

Madhusudan Bhandary is an Associate Professor in the Department of Mathematics. Email: bhandary_madhusudan@colstate.edu. Koji Fujiwara is a graduate student in the Department of Statistics. Email: koji.fujiwara@ndsu.edu.

the equality of two intraclass correlation coefficients based on two independent multinormal samples. For several populations and unequal family sizes, Bhandary and Alam (2000) proposed the Likelihood ratio and large sample ANOVA tests for the equality of several intraclass correlation coefficients based on several independent multinormal samples. Donner and Zou (2002) proposed asymptotic test for the equality of dependent intraclass correlation coefficients under unequal family sizes.

None of the above authors, however, derived any confidence interval estimator for intraclass correlation coefficients under unequal family sizes. In this article, confidence interval estimators for intraclass correlation coefficients are considered based on a single multinormal sample under unequal family sizes, and conditional analyses - assuming family sizes are fixed - though unequal.

It could be of interest to estimate the blood pressure or cholesterol or lung capacity for families in American races. Therefore, an interval estimator for the intraclass correlation coefficient under unequal family sizes must be developed. To address this need, this paper proposes two confidence interval estimators for the intraclass correlation coefficient under unequal family sizes, and these interval estimators are compared using simulation techniques.

Methodology

Proposed Confidence Intervals: Interval Based on the Standard Normal Distribution

Consider a random sample of k families. Let

$$\underset{\sim}{X}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip_i} \end{pmatrix}$$

be a $p_i x1$ vector of observations from i^{th} family; $i = 1,2,...,k$. The structure of the mean vector and the covariance matrix for the familial data is given by the following (Rao, 1973):

$$\underset{\sim}{\mu}_i = \mu \underset{\sim}{1}_i$$

and

$$\underset{p_i x p_i}{\Sigma_i} = \sigma^2 \begin{pmatrix} 1 & \rho & ... & \rho \\ \rho & 1 & ... & \rho \\ ... & ... & ... & ... \\ \rho & \rho & ... & 1 \end{pmatrix}, \qquad (2.1)$$

where $\underset{\sim}{1}_i$ is a $p_i x1$ vector of 1's, $\mu(-\infty < \mu < \infty)$ is the common mean and $\sigma^2(\sigma^2 > 0)$ is the common variance of members of the family and ρ, which is called the intraclass correlation coefficient, is the coefficient of correlation among the members of the family and $\max_{1 \le i \le k}\left(-\frac{1}{p_i - 1}\right) \le \rho \le 1.$

It is assumed that $\underset{\sim}{x}_i \sim N_{p_i}(\underset{\sim}{\mu}_i, \Sigma_i); i = 1,...,k$, where N_{p_i} represents a p_i-variate normal distribution and $\underset{\sim}{\mu}_i, \Sigma_i's$ are defined in (2.1). Let

$$\text{Let } \underset{\sim}{u}_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ . \\ . \\ . \\ u_{ip_i} \end{pmatrix} = Q \underset{\sim}{x}_i \qquad (2.2)$$

where Q is an orthogonal matrix. Under the orthogonal transformation (2.2), it can be shown that $\underset{\sim}{u}_i \sim N_{p_i}(\underset{\sim}{\mu}_i^*, \Sigma_i^*); i = 1,...,k$, where

$$\underset{\sim}{\mu_i^*}_{p_i x 1} = \begin{pmatrix} \mu \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$$

and

$$\Sigma_i^* = \sigma^2 \begin{pmatrix} \eta_i & 0 & ... & 0 \\ 0 & 1-\rho & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1-\rho \end{pmatrix}$$

and $\eta_i = p_i^{-1}\{1+(p_i-1)\rho\}$. The transformation used on the data from $\underset{\sim}{x}$ to $\underset{\sim}{u}$ above is independent of ρ. Helmert's orthogonal transformation can also be used.

Srivastava (1984) gave an estimator of ρ and σ^2 under unequal family sizes which are good substitute for the maximum likelihood estimator and are given by the following:

$$\hat{\rho} = 1 - \frac{\hat{\gamma}^2}{\hat{\sigma}^2},$$

where

$$\hat{\sigma}^2 = (k-1)^{-1}\sum_{i=1}^{k}(u_{i1}-\hat{\mu})^2 + k^{-1}\hat{\gamma}^2\left(\sum_{i=1}^{k}a_i\right)$$

$$\hat{\gamma}^2 = \frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\sum_{i=1}^{k}(p_i-1)}$$

$$\hat{\mu} = k^{-1}\sum_{i=1}^{k}u_{i1} \qquad (2.3)$$

and $a_i = 1 - p_i^{-1}$.

Srivastava and Katapa (1986) derived the asymptotic distribution of $\hat{\rho}$; they showed: that $\hat{\rho} \sim \mathrm{N}(\rho, Var/k)$ asymptotically, where

$$Var = 2(1-\rho)^2 \left\{(\bar{p}-1)^{-1} + c^2 - 2(1-\rho)(\bar{p}-1)^{-1}k^{-1}\sum_{i=1}^{k}a_i\right\} \qquad (2.4)$$

k = number of families in the sample

$$\bar{p} = k^{-1}\sum_{i=1}^{k}p_i$$

$$c^2 = 1 - 2(1-\rho)^2 k^{-1}\sum_{i=1}^{k}a_i + (1-\rho)^2\left[k^{-1}\sum_{i=1}^{k}a_i + (\bar{p}-1)^{-1}(k^{-1}\sum_{i=1}^{k}a_i)^2\right]$$

and $a_i = 1 - p_i^{-1}$.

Under the above setup, it is observed (using Srivastava & Katapa, 1986) that:

$$Z = \frac{\hat{\rho}-\rho}{\sqrt{\frac{Var}{k}}} \sim N(0,1), \qquad (2.5)$$

asymptotically, where, Var is to be determined from (2.4) and $\hat{\rho}$ is obtained from (2.3).

Using the expression (2.5), it is found that the $100(1-\alpha)\%$ confidence interval for ρ is

$$\hat{\rho} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{Var}{k}} \qquad (2.6)$$

Interval Based on the χ^2 Distribution

It can be shown, by using the distribution of $\underset{\sim}{u_i}$ given by (2.2):

$$\frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\sigma^2(1-\rho)} \sim \chi^2_{\sum_{i=1}^{k}(p_i-1)}, \qquad (2.7)$$

where χ_n^2 denotes the Chi-square distribution with n degrees of freedom. Using (2.7), a $100(1-\alpha)\%$ confidence interval for ρ can be found as follows:

$$1 - \frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\chi^2_{1-\frac{\alpha}{2};n}.\hat{\sigma}^2} < \rho < 1 - \frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2}{\chi^2_{\frac{\alpha}{2};n}.\hat{\sigma}^2} \qquad (2.8)$$

where, $n = \sum_{i=1}^{k} (p_i - 1)$ and $\chi^2_{\alpha;n}$ denotes the upper 100α % point of a Chi-square distribution with n degrees of freedom and $\hat{\sigma}^2$ can be obtained from (2.3).

Data Simulation

Multivariate normal random vectors were generated using an R program in order to evaluate the average lengths and coverage probability of the intervals given by (2.6) and (2.8). Fifteen and 30 vectors of family data were created for the population. The family size distribution was truncated to maintain the family size at a minimum of two siblings and a maximum of 15 siblings. The previous research in simulating family sizes (Rosner, et al., 1977; and Srivastava & Keen, 1988) determined the parameter setting for FORTRAN IMSL negative binomial subroutine with a mean = 2.86 and a success probability = 0.483.

Given this, the mean was set to equal 2.86 and theta was set to equal 41.2552. All parameters were set the same except for the value of ρ which took values from 0.1 to 0.9 at increments of 0.1.The R program produced 3,000 estimates of ρ along with the coverage probability and the confidence intervals given by the formulae (2.6) and (2.8) for each particular value of the population parameter ρ. The average length and coverage probability of each interval at α=0.01, 0.05 and 0.10 were noted. Results are shown in Table1.

Table 1: Coverage Probability and Length for the Confidence Interval

			Coverage Probability		Length	
rho	k	alpha	Z	Chi-square	Z	Chi-square
0.1	15	0.01	1.00000	0.99933	1.04368	1.06377
0.2	15	0.01	0.98533	0.99733	1.12548	1.05273
0.3	15	0.01	0.99233	0.99300	1.08944	0.90430
0.4	15	0.01	0.98400	0.98167	1.09402	1.01612
0.5	15	0.01	0.98333	0.95133	0.95383	0.75022
0.1	15	0.05	0.92500	0.98800	0.94959	1.16835
0.2	15	0.05	0.96433	0.97367	0.87928	0.81043
0.3	15	0.05	0.97033	0.94933	0.83507	0.67550
0.4	15	0.05	0.95800	0.92067	0.83382	0.73789
0.2	15	0.10	0.95233	0.92067	0.71398	0.57282
0.3	15	0.10	0.95433	0.91067	0.69647	0.55067
0.4	15	0.10	0.95200	0.83500	0.65522	0.46074
0.1	30	0.01	1.00000	0.99967	0.79989	0.73312
0.2	30	0.01	0.99767	0.99667	0.82135	0.68646
0.3	30	0.01	0.99533	0.98833	0.80516	0.63780
0.4	30	0.01	0.99433	0.98167	0.76184	0.59005
0.5	30	0.01	0.99400	0.96867	0.67756	0.49657
0.6	30	0.01	0.99167	0.94500	0.57519	0.40045
0.7	30	0.01	0.98967	0.91200	0.44465	0.27996
0.1	30	0.05	0.96900	0.98567	0.64870	0.63591
0.2	30	0.05	0.97867	0.97333	0.66055	0.59177
0.3	30	0.05	0.98000	0.94533	0.61955	0.48249
0.4	30	0.05	0.97600	0.91800	0.57706	0.43160
0.1	30	0.10	0.96267	0.97633	0.53021	0.51577
0.2	30	0.10	0.96100	0.93867	0.54511	0.46834
0.3	30	0.10	0.96133	0.87933	0.51242	0.38011
0.4	30	0.10	0.94400	0.86567	0.49921	0.39224

The interval based on the χ^2 distribution given by (2.8) showed consistently better results in terms of shorter average interval length compared to the interval based on the standard normal distribution given by (2.6), especially for larger sample sizes for various intraclass correlation coefficient values. The average lengths and coverage probability of both intervals are presented in Table 1. The interval based on the χ^2 distribution is recommended on the basis of shorter average interval length. The coverage probability of the interval based on the χ^2 distribution is competitive with the coverage probability of the interval based on the standard normal distribution.

Real Data Example

Two intervals using real life data collected from Srivastava and Katapa (1986) were compared. The real life data presented in Srivastava and Katapa (1986) is shown in Table 2.

Table 2: Values of Pattern Intensity on Soles of Feet in Fourteen Families

Sample	Family #	Mother	Father	Siblings
A	12	2	4	2, 4
A	10	5	4	4, 5, 4
A	9	5	5	5, 6
A	1	2	3	2, 2
A	4	2	4	2, 2, 2, 2, 2
A	5	6	7	6, 6
A	8	3	7	2, 4, 7, 4, 4, 7, 8
A	3	2	3	2, 2, 2
A	6	4	3	4, 3, 3
A	14	2	3	2, 2, 2
A	7	4	3	2, 2, 3, 6, 3, 5, 4
A	2	2	3	2, 3
A	11	5	6	5, 3, 4, 4
A	13	6	3	4, 3, 3, 3

The data is first transformed by multiplying each observation vector by Helmert's orthogonal matrix Q, where

$$\underset{p_i \times p_i}{Q} = \begin{bmatrix} \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \cdots & \frac{1}{\sqrt{p_i}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \cdots & -\frac{(p_i-1)}{\sqrt{p_i(p_i-1)}} \end{bmatrix},$$

which results in transformed vectors $\underset{\sim}{u}_i$ for $i = 1,2,...,k$, here, k = 14. Srivastava's formula given by (2.3) is used to compute the intraclass correlation coefficient and variance. The computed values of the intraclass correlation coefficient and variance are $\hat{\rho} = 0.8118$ and $\hat{\sigma}^2 = 8.8578$. Using formulae (2.6) and (2.8) to obtain the lengths of the 95%, 99% and 90% confidence intervals for the intraclass correlation coefficient results in the following:

- Length of 90% confidence interval based on Z-distribution = 0.26519
- Length of 90% confidence interval based on χ^2 - distribution = 0.16096
- Length of 95% confidence interval based on Z-distribution = 0.31599
- Length of 95% confidence interval based on χ^2 - distribution = 0.19644
- Length of 99% confidence interval based on Z-distribution = 0.41528
- Length of 99% confidence interval based on χ^2 - distribution = 0.27388

It is observed that the length of the 95%, 99% and 90% confidence intervals based on the χ^2 distribution(using formula 2.8) is shorter than the length of the 95%, 99% and 90% confidence intervals respectively based on standard normal distribution (using formula 2.6).

References

Bhandary, M., & Alam, M. K. (2000). Test for the equality of intraclass correlation coefficients under unequal family sizes for several populations. *Communications in Statistics-Theory and Methods, 29*(*4*), 755-768.

Donner, A., & Bull, S. (1983). Inferences concerning a common intraclass correlation coefficient. *Biometrics, 39*, 771-775.

Donner, A., & Zou, G. (2002). Testing the equality of dependent intraclass correlation coefficients. *The Statistician*, *51*(*3*), 367-379.

Gokhale, D. V., & SenGupta, A. (1986). Optimal tests for the correlation coefficient in a symmetric multivariate normal population. *Journal of Statistical Planning Inference*, *14*, 263-268.

Huang, W., & Sinha, B. K. (1993). On optimum invariant tests of equality of intraclass correlation coefficients. *Annals of the Institute of Statistical Mathematics*, *45*(*3*), 579-597.

Konishi, S. (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics*, *37*, 87-94.

Konishi, S., & Gupta, A. K. (1989). Testing the equality of several intraclass correlation coefficients. *Journal of Statistical Planning Inference*, *21*, 93-105.

Rao, C. R. (1973). *Linear statistical inference and its applications*. NY: Wiley.

Rosner, B., Donner, A., & Hennekens, C. H. (1977). Estimation of intraclass correlation from familial data. *Applied Statistics*, *26*, 179-187.

Rosner, B., Donner, A., & Hennekens, C. H.(1979). Significance testing of interclass correlations from familial data. *Biometrics*, *35*, 461-471.

SenGupta, A. (1988). On loss of power under additional information – an example. *Scandinavian Journal of Statistics*, *15*, 25-31.

Scheffe, H. (1959). *The analysis of variance*. NY: Wiley.

Srivastava, M.S. (1984). Estimation of interclass correlations in familial data. *Biometrika*, *71*, 177-185.

Srivastava, M. S., & Katapa, R. S. (1986). Comparison of estimators of interclass and intraclass correlations from familial data. *Canadian Journal of Statistics*, *14*, 29-42.

Srivastava, M. S., & Keen, K. J. (1988). Estimation of the interclass correlation coefficient. *Biometrika*, *75*, 731-739.

Velu, R., & Rao, M. B. (1990). Estimation of parent-offspring correlation. *Biometrika*, *77*(*3*), 557-562.

Young, D., & Bhandary, M. (1998). Test for the equality of intraclass correlation coefficients under unequal family sizes. *Biometrics*, *54*(*4*), 1363-1373.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 526-533

1538 – 9472/09/$95.00

Approximate Bayesian Confidence Intervals for The Mean of a Gaussian Distribution Versus Bayesian Models

Vincent A. R. Camara
University of South Florida

This study obtained and compared confidence intervals for the mean of a Gaussian distribution. Considering the square error and the Higgins-Tsokos loss functions, approximate Bayesian confidence intervals for the mean of a normal population are derived. Using normal data and SAS software, the obtained approximate Bayesian confidence intervals were compared to a published Bayesian model. Whereas the published Bayesian method is sensitive to the choice of the hyper-parameters and does not always yield the best confidence intervals, it is shown that the proposed approximate Bayesian approach relies only on the observations and often performs better.

Key words: Estimation; loss functions; confidence intervals, statistical analysis.

Introduction

A significant amount of research in Bayesian analysis and modeling has been published during the last twenty-five years. Bayesian analysis implies the exploitation of suitable prior information and the choice of a loss function in association with Bayes' Theorem. It rests on the notion that a parameter within a model is not merely an unknown quantity, but behaves as a random variable that follows some distribution. In the area of life testing, it is realistic to assume that a life parameter is stochastically dynamic. This assertion is supported by the fact that the complexity of electronic and structural systems is likely to cause undetected component interactions resulting in an unpredictable fluctuation of the life parameter.

Vincent A. R. Camara earned a Ph.D. in Mathematics/Statistics. His research interests include the theory and applications of Bayesian and empirical Bayes analyses with emphasis on the computational aspect of modeling. This research paper has been sponsored by the Research Center for Bayesian Applications, Inc
E-mail: gvcamara@ij.net

Although no specific analytical procedure exists which identifies the appropriate loss function to be used, the most commonly used is the square error loss function. One reason for selecting this loss function is due to its analytical tractability in Bayesian modeling and analysis.

The square error loss function places a small weight on estimates near the parameter's true value and proportionately more weight on extreme deviations from the true value. The square error loss is defined as follows:

$$L_{SE}(\overset{\Lambda}{\boldsymbol{\theta}},\boldsymbol{\theta}) = \left(\overset{\Lambda}{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^2.$$

This study considers a widely used and useful underlying model, the normal underlying model, which is characterized by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};$$

$$-\infty \prec x \prec \infty, -\infty \prec \mu \prec \infty, \sigma \succ 0. \quad (1)$$

Employing the square error loss function along with a normal prior, Fogel (1991) obtained the following Bayesian confidence interval for the mean of the normal probability density function:

$$L_B = \frac{\mu_1\sigma^2/n + +\bar{x}\tau^2}{\tau^2+\sigma^2/n} - Z_{\alpha/2}\frac{\tau\sigma/\sqrt{n}}{\sqrt{\tau^2+\sigma^2/n}} \tag{2}$$

$$U_B = \frac{\mu_1\sigma^2/n + +\bar{x}\tau^2}{\tau^2+\sigma^2/n} + Z_{\alpha/2}\frac{\tau\sigma/\sqrt{n}}{\sqrt{\tau^2+\sigma^2/n}} \tag{3}$$

where the mean and variance of the selected normal prior are respectively denoted by μ_1 and τ^2.

This study employs the square error and the Higgins-Tsokos loss functions to derive approximate Bayesian confidence intervals for the normal population mean. Obtained confidence bounds are then compared with their Bayesian counterparts corresponding to (3).

Methodology

Considering the normal density function (2), to derive approximate Bayesian confidence intervals for the mean of a normal distribution, results obtained on approximate Bayesian confidence intervals for the variance of a Gaussian distribution are used (Camara, 2003). The loss functions used are the square error loss function (1), and the Higgins-Tsokos loss function.

The Higgins-Tsokos loss function places a heavy penalty on extreme over- or under-estimation. That is, it places an exponential weight on extreme errors. The Higgins-Tsokos loss function is defined as follows:

$$L_{HT}(\hat{\theta},\theta) = \frac{f_1 e^{f_2(\hat{\theta}-\theta)} + f_2 e^{-f_1(\hat{\theta}-\theta)}}{f_1+f_2} - 1,$$

$$f_1, f_2 \succ 0. \tag{4}$$

The use of these loss functions (1) and (4), along with suitable approximations of the Pareto prior, led to the following approximate Bayesian confidence bounds for the variance of a normal population (Camara, 2003). For the square error loss function:

$$L_{\sigma^2(SE)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-2-2\ln(\alpha/2)}$$

$$U_{\sigma^2(SE)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-2-2\ln(1-\alpha/2)} \tag{5}$$

For the Higgins-Tsokos loss function:

$$U_{\sigma^2(HT)} = \frac{1}{\dfrac{n-1-2Ln(1-\alpha/2)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} - G(x)}$$

$$f_1 \prec \frac{(x-\mu)^2}{2}, \tag{6}$$

where

$$G(x) = \frac{1}{f_1+f_2} Ln\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + f_2}{\sum_{i=1}^{n}(x_i-\bar{x})^2 - f_1}\right). \tag{7}$$

Using the above approximate Bayesian confidence intervals for a normal population variance (5) (6) along with

$$\sigma^2 = E(X^2) - \mu^2, \tag{8}$$

the following approximate Bayesian confidence intervals for the mean of a normal population can easily be derived for a strictly positive mean.

The approximate Bayesian confidence interval for the normal population mean corresponding to the square error loss is:

$$L_{\mu(SE)} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} + \bar{x}^2 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 2 - 2\ln(1-\alpha/2)} \right)^{0.5}$$

$$U_{\mu(SE)} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} + \bar{x}^2 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 2 - 2\ln(\alpha/2)} \right)^{0.5} \quad (9)$$

The approximate Bayesian confidence interval for the normal population mean corresponding to the Higgins-Tsokos loss function is:

$$U_{\mu(HT)} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} + \bar{x}^2 - \frac{1}{H_2(x)} \right)^{0.5} \quad (10)$$

where

$$H_1(x) = \frac{n - 1 - 2Ln(1-\alpha/2)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} - G(x) \quad (11)$$

$$H_2(x) = \frac{n - 1 - 2Ln(\alpha/2)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} - G(x) \quad (12)$$

$$L_{\mu(HT)} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} + \bar{x}^2 - \frac{1}{H_1(x)} \right)^{0.5}$$

With (9),(10), (11), (12) and a change of variable, approximate Bayesian Confidence intervals are easily obtained when $\mu \le 0$.

Results

To compare the Bayesian model (3) with the approximate Bayesian models (9 & 10), samples obtained from normally distributed populations (Examples 1, 2, 3, 4, 7) as well as approximately normal populations (Examples 5, 6) were considered. SAS software was employed to obtain the normal population parameters corresponding to each sample data set. For the Higgins-Tsokos loss function, f1 = 1 and f2 = 1 were considered.

Example 1

Data Set: 24, 28, 22, 25, 24, 22, 29, 26, 25, 28, 19, 29 (Mann, 1998, p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 25.083, \sigma = 3.1176),$$

$$\bar{x} = 25.08333,\ s^2 = 9.719696.$$

Table 1a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding Data Set 1

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	25.0683-25.1311	25.0730-25.1158
90	25.0661-25.1437	25.0683-25.1311
95	25.0650-25.1543	25.0661-25.1437
99	25.0641-25.1734	25.0643-25.1660

Table 1b: Bayesian Confidence Intervals for the Population Mean Corresponding Data Set 1

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	13.8971-15.6097	23.9353-26.2300
90	13.6496-15.8572	23.6037-26.5617
95	13.4422-16.0646	23.3258-26.8395
99	13.0275-16.4793	22.7701-27.3953

Example 2

Data Set: 13, 11, 9, 12, 8, 10, 5, 10, 9, 12, 13 (Mann, 1998, p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 10.182, \sigma = 2.4008),$$

$$\bar{x} = 10.181812,\ s^2 = 5.763636.$$

Table 2a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 2

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	10.1575-10.2565	10.1652-10.2330
90	10.1538-10.2756	10.1575-10.2565
95	10.1520-10.2914	10.1538-10.2756
99	10.1506-10.3194	10.1506-10.3194

Table 2b: Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 2

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	6.6182-8.1193	9.3349-11.1832
90	6.4013-8.3363	9.0678-11.4503
95	6.2195-8.5180	8.8440-11.6741
99	5.8560-8.8816	8.3964-12.1217

Example 3

Data Set: 16, 14, 11, 19, 14, 17, 13, 16, 17, 18, 19, 12 (Mann, 1998, p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 15.5, \sigma = 2.6799),$$

$$\bar{x} = 15.5,\ s^2 = 7.181818.$$

Table 3a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 3

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	15.4820-15.5570	15.4877-15.5388
90	15.4794-15.5721	15.4820-15.5570
95	15.4781-15.5847	15.4794-15.5721
99	15.4770-15.6075	15.4773-15.5986

Table 3b: Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 3

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	9.6623-11.2287	14.5692-16.5438
90	9.4359-11.4551	14.2839-16.8292
95	9.2462-11.6448	14.0447-17.0683
99	8,8668-12.0242	13.5665-17.5465

Example 4

Data Set: 27, 31, 25, 33, 21, 35, 30, 26, 25, 31, 33, 30, 28 (Mann, 1998, p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 28.846, \sigma = 3.9549),$$

$$\bar{x} = 28.846153,\ s^2 = 15.641025.$$

Table 4a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 4

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	28.8270-28.9087	28.8330-28.8884
90	28.8242-28.9256	28.8270-28.9087
95	28.8228-28.9400	28.8242-28.9256
99	28.8217-28.9663	28.8220-28.9560

Table 5a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 5

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	43.5794-43.6703	43.5858-43.6169
90	43.5764-43.6902	43.5794-43.6703
95	43.5749-43.7074	43.5764-43.6902
99	43.5738-43.7395	43.5741-43.7268

Table 4b: Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 4

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	13.2394-15.1312	27.4048-30.1961
90	12.9659-15.4047	27.0014-30.5995
95	12.7369-15.6337	26.6634-30.9375
99	12.2787-16.0919	25.9873-31.6135

Table 5b: Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 5

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	14.8305-16.9204	41.4441-45.0272
90	14.5285-17.2225	40.9263-45.5450
95	14.2754-17.4756	40.4924-45.9789
99	13.7692-17.9817	39.6246-46.8467

Example 5

Data Set: 52, 33, 42, 44, 41, 50, 44, 51, 45, 38,37,40,44, 50, 43 (McClave & Sincich, p. 301).

Normal population distribution obtained with SAS:

$$N(\mu = 43.6, \sigma = 5.4746),$$
$$\bar{x} = 43.6,\ s^2 = 29.971428.$$

Example 6

Data Set: 52, 43, 47, 56, 62, 53, 61, 50, 56, 52, 53, 60, 50, 48, 60, 5543 (McClave & Sincich, p. 301).

Normal population distribution obtained with SAS:

$$N(\mu = 53.625, \sigma = 5.4145)$$
$$\bar{x} = 53.625,\ s^2 = 29.316666.$$

Table 6a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 6

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	53.6098-53.6779	53.6145-53.6602
90	53.6076-53.6932	53.6098-53.6779
95	53.6065-53.7064	53.6076-53.6932
99	53.6056-53.7315	53.6058-53.7216

Table 6b: Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 6

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	19.1978-21.2568	51.3930-54.8269
90	18.9002-21.5544	50.8967-55.3232
95	18.6508-21.8038	50.4808-55.7391
99	18.1521-22.3024	49.6492-56.5707

Table 7a: Approximate Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 7

C.L. %	Approximate Bayesian Bounds (SE)	Approximate Bayesian Bounds (HT)
80	82.7072-83.4808	82.7539-83.2572
90	82.6856-83.6884	82.7072-83.4808
95	82.6751-83.8815	82.6856-83.6884
99	82.6669-84.2823	82.6690-83.7173

Table 7b: Bayesian Confidence Intervals for the Population Mean Corresponding to Data Set 7

C.L. %	Bayesian C. I. I Bayesian Bounds $\mu_1 = 2, \tau = 1$	Bayesian C. I. II Bayesian Bounds $\mu_1 = 25, \tau = 10$
80	3.2940-5.8132	63.0810-75.4828
90	2.9299-6.17740	61.2886-77.2752
95	2.6248-6.4824	59.7868-78.7770
99	2.0147-7.0926	56.7833-81.7806

Example 7

Data Set: 50, 65, 100, 45, 111, 32, 45, 28, 60, 66, 114, 134, 150, 120, 77, 108, 112, 113, 80, 77, 69, 91, 116, 122, 37, 51, 53, 131, 49, 69, 66, 46, 131, 103, 84, 78 (SAS Data).

Normal population distribution obtained with SAS:

$$N(\mu = 82.861, \sigma = 33.226)$$

$$\bar{x} = 82.8611,\ s^2 = 1103.951587$$

All seven Examples show that the proposed approximate Bayesian confidence intervals contain the population mean. The Bayesian model, however, does not always contain the population mean.

Conclusion

In this study, approximate Bayesian confidence intervals for the mean of a normal population under two different loss functions were derived and compared with a published Bayesian model (Fogel, 1991). The loss functions employed were the square error and the Higgins-Tsokos

loss functions. The following conclusions are based on results obtained:

1. The Bayesian model (3) used to construct confidence intervals for the mean of a normal population does not always yield the best coverage accuracy. Each of the obtained approximate Bayesian confidence intervals contains the population mean and performs better than its Bayesian counterparts.
2. Bayesian models are generally sensitive to the choice of hyper-parameters. Some values arbitrarily assigned to the hyper-parameters may lead to a very poor estimation of the parameter(s) under study. In this study some values assigned to the hyper-parameters led to confidence intervals that do not contain the normal population mean.
3. Contrary to the Bayesian model (3), which uses the Z-table, both the approach employed in this study and our approximate Bayesian models rely only on observations.
4. With the proposed approach, approximate Bayesian confidence intervals for a normal population mean are easily obtained for any level of significance..
5. The approximate Bayesian approach under the popular square error loss function does not always yield the best approximate Bayesian results: The Higgins-Tsokos loss function performs better in the examples presented.

References

Bhattacharya, S. K. (1967) Bayesian approach to life testing and reliability estimation. *Journal of the American Statistical Association*, *62*, 48-62.

Britney, R. R., & Winkler, R. L. (1968). *Bayesian III point estimation under various loss functions*. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 356-364.

Camara, V. A. R. (2002). *Approximate Bayesian confidence intervals for the variance of a Gaussian distribution***.** Proceedings of the American Statistical Association, Statistical Computing Section. NY: American Statistical Association.

Camara, V. A. R. (2003). Approximate Bayesian confidence intervals for the variance of a Gaussian Distribution**.** *Journal of Modern Applied Statistical Methods*, *2*(*2*), 350-358.

Camara, V. A. R., & Tsokos, C. P. (1996). *Effect of loss functions on Bayesian reliability analysis*. Proceedings of the International Conference on Nonlinear Problems in Aviation and Aerospace, 75-90.

Camara, V. A. R., & Tsokos, C. P. (1998). *Bayesian reliability modeling with applications*. NY: UMI Publishing Company.

Camara, V. A. R., & Tsokos, C. P. (1999). Bayesian estimate of a parameter and choice of the loss function. *Nonlinear Studies Journal*, VOL 6 No 1 pp. 55-64

Camara, V. A. R., & Tsokos, C. P. (2001). Sensitivity Behavior of Bayesian Reliability Analysis for different Loss Functions, *International Journal of Applied Mathematics*, VOL 6 pp . 35-38.

Camara, V. A. R., & Tsokos, C. P. (1998). The effect of loss functions on empirical Bayes reliability analysis. *Journal of Engineering Problems*, VOL 4 pp 539-560

Canfield, R. V. (1970). A Bayesian approach to reliability estimation using a loss function, *IEEE Trans. Reliability*, *R-19*(*1*), 13-16.

Drake, A. W. (1966). *Bayesian statistics for the reliability engineer*. Proceedings from the Annual Symposium on Reliability, 315-320.

Fogel, M. (1991) Bayesian Confidence Interval. The Statistics Problem Solver, 502-505. Research& Education Association.

Harris, B. (1976). A survey of statistical methods in system reliability using Bernoulli sampling of components. In *Proceedings of the conference on the theory and applications of Reliability with emphasis on Bayesian and Nonparametric Methods*. NY: Academic Press.

Higgins, J. J., & Tsokos, C. P. (1976). Comparison of Bayes estimates of failure intensity for fitted priors of life data. In *Proceedings of the Conference on the Theory an Applications of Reliability with Emphasis on Bayesian and Nonparametric Methods*. NY: Academic Press.

Higgins, J. J., & Tsokos, C. P. (1976). On the behavior of some quantities used in Bayesian reliability demonstration tests, IEEE Trans. *Reliability*, *R-25*(*4*), 261-264.

Higgins, J. J., & Tsokos, C. P. (1980). A study of the effect of the loss function on Bayes estimates of failure intensity, MTBF, and reliability. *Applied Mathematics and Computation*, *6*, 145-166.

Mann, P. S. (1998). *Introductory statistics* (*3rd Ed.*).John Wiley & Sons, Inc, New York.

McClave, J. T., & Sincich, T. A. (1997). *First course in statistics*, (*6th Ed.*). NY: Prentice Hall.

Schafer, R. E., et al. (1970). *Bayesian reliability demonstration, phase I: data for the a priori distribution*. Rome Air Development Center, Griffis AFBNY RADC-TR-69-389.

Schafer, R. E., et al. (1971). *Bayesian reliability, phase II: Development of a priori distribution*. Rome Air Development Center, Griffis AFR, NY RADC-YR-71-209.

Schafer, R. E., et al. (1973). *Bayesian reliability demonstration phase III: Development of test plans*. Rome Air development Center, Griffs AFB, NY RADC-TR-73-39.

Shafer, R. E., & Feduccia, A. J. (1972). Prior distribution fitted to observed reliability data. *IEEE Trans. Reliability*, *R-21*(*3*), 148-154

Tsokos, C. P., & Shimi, S. (Eds). (1976). *Proceedings of the Conference on the theory and applications of reliability with emphasis on Bayesian and nonparametric methods, Methods, Vols. I, II*. NY: Academic Press.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 534-546

1538 – 9472/09/$95.00

On Type-II Progressively Hybrid Censoring

Debasis Kundu
Indian Institute of Technology, Kanpur, India

Avijit Joarder
Reserve Bank of India, Mumbai, India

Hare Krishna
C.C.S. University, Meerut, India

The progressive Type-II censoring scheme has become quite popular. A drawback of a progressive censoring scheme is that the length of the experiment can be very large if the items are highly reliable. Recently, Kundu and Joarder (2006) introduced the Type-II progressively hybrid censored scheme and analyzed the data assuming that the lifetimes of the items are exponentially distributed. This article presents the analysis of Type-II progressively hybrid censored data when the lifetime distributions of the items follow Weibull distributions. Maximum likelihood estimators and approximate maximum likelihood estimators are developed for estimating the unknown parameters. Asymptotic confidence intervals based on maximum likelihood estimators and approximate maximum likelihood estimators are proposed. Different methods are compared using Monte Carlo simulations and one real data set is analyzed.

Key words: Maximum likelihood estimators; approximate maximum likelihood estimators; Type-I censoring; Type-II censoring; Monte Carlo simulation.

Introduction

The Type-II progressive censoring scheme has become very popular. It can be described as follows: consider n units in a study and suppose $m < n$ is fixed before the experiment, in addition, m other integers, $R_1,..,R_m$ are also fixed so that $R_1 + ... + R_m + m = n$. At the time of the first failure, for example, $Y_{1:m:n}$, R_1 of the remaining units are randomly removed. Similarly, at the time of the second failure, for example, $Y_{2:m:n}$, R_2 of the remaining units are randomly removed and so on. Finally, at the time of the $m-th$ failure, $Y_{m:m:n}$, the remaining R_m units are removed. Extensive work has been conducted on this particular scheme during the last ten years; see Balakrishnan and Aggarwala (2000) and Balakrishnan (2007).

Unfortunately the major problem with the Type-II progressive censoring scheme is that the time length of the experiment can be very large. Due to this problem, Kundu and Joarder (2006) introduced a new censoring scheme named Type-II Progressively Hybrid Censoring, which ensures that the length of the experiment cannot exceed a pre-specified time point T. The detailed description and advantages of the Type-II progressively hybrid censoring is presented in Kundu and Joarder (2006) (see also Childs, Chandrasekar & Balakrishnan, 2007); in both publications the authors assumed the lifetime distributions of the items to be exponential.

Because the exponential distribution has limitations, this article considers the Type-II progressively hybrid censored lifetime data, when the lifetime follows a two-parameter Weibull distribution. Maximum likelihood estimators (MLEs) of the unknown parameters are provided and it was observed that the MLEs cannot be obtained in explicit forms. MLEs can

Debasis Kundu is a Professor of the Department of Mathematics and Statistics, Kanpur IIT. E-mail: kundu@iitk.ac.in. Avijit Joarder is an Assistant Adviser of Department of Statistics and Information Management, RBI. Please note that the views in this article are A. Joarder's personal views and not those of the RBI. E-mail: ajoarder@rbi.org.in. Hare Krishna is the Head of the Department of Statistics, CCS University. E-mail: hkrishnastats@yahoo.com.

be obtained by solving a non-linear equation and a simple iterative scheme is proposed to solve the non-linear equation. Approximate maximum likelihood estimators (AMLEs), which have explicit expressions are also suggested. It is not possible to compute the exact distributions of the MLEs, so the asymptotic distribution is used to construct confidence intervals. Monte Carlo simulations are used to compare different methods and one data analysis is performed for illustrative purposes.

Type-II Progressively Hybrid Censoring Scheme Models

If it is assumed that the lifetime random variable Y has a Weibull distribution with shape and scale parameters α and λ respectively, then the probability density function (PDF) of Y is

$$f_Y(y;\alpha,\gamma)=\frac{\alpha}{\lambda}\left(\frac{y}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{y}{\lambda}\right)^{\alpha}}; \quad y>0, \quad (1)$$

where $\alpha>0, \lambda>0$ are the natural parameter space. If the random variable Y has the density function (1), then $X=\ln Y$ has the extreme value distribution with the PDF

$$f_X(x;\mu,\sigma)=\frac{1}{\sigma}e^{\left(\frac{x-\mu}{\sigma}-e^{\frac{x-\mu}{\sigma}}\right)}; \quad -\infty<x<\infty, \quad (2)$$

where $\mu=\ln\lambda, \sigma=1/\alpha$. The density function as described by (2) is known as the density function of an extreme value distribution with location and scale parameters μ and σ respectively. Models (1) and (2) are equivalent models in the sense that the procedure developed under one model can be easily used for the other model. Although, they are equivalent models, (2) can be the easier with which to work compared to model (1), because in model (2) the two parameters μ and σ appear as location and scale parameters. For $\mu=0$ and $\sigma=1$, model (2) is known as the standard extreme value distribution and has the following PDF

$$f_Z(z;0,1)=e^{\left(z-e^{z}\right)}; \quad -\infty<z<\infty. \quad (3)$$

Type-II Progressively Hybrid Censoring Scheme Data

Under the Type-II progressively hybrid censoring scheme, it is assumed that n identical items are put on a test and the lifetime distributions of the n items are denoted by $Y_1,..,Y_n$. The integer $m<n$ is pre-fixed, $R_1,..,R_m$ are m pre-fixed integers satisfying $R_1+.....+R_m+m=n$, and T is a pre-fixed time point. At the time of the first failure $Y_{1:m:n}$, R_1 of the remaining units are randomly removed. Similarly, at the time of the second failure $Y_{2:m:n}$, R_2 of the remaining units are removed and so on. If the $m-th$ failure $Y_{m:m:n}$ occurs before time T, the experiment stops at time point $Y_{m:m:n}$. If, however, the m-th failure does not occur before time point T and only J failures occur before T (where $0\le J<m$), then at time T all remaining R_J^* units are removed and the experiment terminates. Note that $R_J^*=n-(R_1+...+R_J)-J$. The two cases are denoted as Case I and Case II respectively and this is called the censoring scheme as the Type-II progressively hybrid censoring scheme (Kundu and Joarder, 2006).

In the presence of the Type-II progressively hybrid censoring scheme, one of the following is observed

Case I:

$$\{Y_{1:m:n},...,Y_{m:m:n}\};\ \text{if } Y_{m:m:n}<T, \quad (4)$$

or

Case II:

$$\{Y_{1:m:n},...,Y_{J:m:n}\};\ \text{if } Y_{J:m:n}<T<Y_{J+1:m:n}. \quad (5)$$

For Case II, although $Y_{J+1:m:n}$ is not observed, but $Y_{J:m:n}<T<Y_{J+1:m:n}$ means that the $J-th$ failure took place before T and no failure took place between $Y_{J:m:n}$ and T (i.e., $Y_{J+1:m:n},..,Y_{m:m:n}$) are not observed.

The conventional Type-I progressive censoring scheme needs the pre-specification of $R_1,..,R_m$ and also $T_1,...,T_m$ (see Cohen 1963, 1966 for details). The choices of $T_1,..,T_m$ are not

trivial. For the conventional Type-II progressive censoring scheme the experimental time is unbounded. In the proposed censoring scheme the choice of T depends on how much maximum experimental time the experimenter can afford to continue and also the experimental time is bounded.

Maximum Likelihood Estimators (MLEs)

Based on the observed data, the likelihood function for Case I is

$$l(\alpha,\lambda)=K_1\left(\frac{\alpha}{\lambda}\right)^m\prod_{i=1}^{m}\left(\frac{y_{i:m:n}}{\lambda}\right)^{\alpha-1}e^{-\left[\sum_{i=1}^{m}(1+R_i)\left(\frac{y_{i:m:n}}{\lambda}\right)^{\alpha}\right]}, \tag{6}$$

and for Case II, the MLE is

$$l(\alpha,\lambda)=K_2\left(\frac{\alpha}{\lambda}\right)^J\prod_{i=1}^{J}\left(\frac{y_{i:m:n}}{\lambda}\right)^{\alpha-1}e^{-\left[\sum_{i=1}^{J}(1+R_i)\left(\frac{y_{i:m:n}}{\lambda}\right)^{\alpha}+R_J^*\left(\frac{T}{\lambda}\right)^{\alpha}\right]} \quad \text{if } J>0, \tag{7}$$

$$=e^{-n\left(\frac{T}{\lambda}\right)^{\alpha}}, \quad \text{if } J=0 \tag{8}$$

where

$$K_1=\prod_{i=1}^{m}\left[n-\sum_{k=1}^{i-1}(1+R_k)\right]$$

and

$$K_2=\prod_{i=1}^{J}\left[n-\sum_{k=1}^{i-1}(1+R_k)\right],$$

both are constant.

The logarithm of (6) and (7), can be written without the constant terms as

$$L(\alpha,\lambda)=d(\ln\alpha-\ln\lambda)+(\alpha-1)\left[\sum_{i=1}^{d}\ln y_{i:m:n}-d\ln\lambda\right]-\frac{1}{\lambda^{\alpha}}W(\alpha). \tag{9}$$

Here $d=m,\ W(\alpha)=\sum_{i=1}^{m}(1+R_i)y_{i:m:n}^{\alpha}$ and $d=J,\ W(\alpha)=\sum_{i=1}^{J}(1+R_i)y_{i:m:n}^{\alpha}+R_J^*T^{\alpha}$ for Case-I and Case-II respectively. It is assumed that $d>0$, otherwise the MLEs do not exist.

Taking derivatives with respect to α and λ of (9) and equating them to zero results in

$$\frac{\partial L(\alpha,\lambda)}{\partial\lambda}=-\frac{d\alpha}{\lambda}+\frac{\alpha}{\lambda^{\alpha+1}}W(\alpha)=0, \tag{10}$$

$$\frac{\partial L(\alpha,\lambda)}{\partial\alpha}=\frac{d}{\alpha}+\sum_{i=1}^{d}\ln y_{i:m:n}-d\ln\lambda-\frac{1}{\lambda^{\alpha}}V(\alpha)+\frac{1}{\lambda^{\alpha}}W(\alpha)\ln\lambda=0. \tag{11}$$

Here, $V(\alpha)=\sum_{i=1}^{m}(1+R_i)y_{i:m:n}^{\alpha}\ln y_{i:m:n}$ and $V(\alpha)=\sum_{i=1}^{J}(1+R_i)y_{i:m:n}^{\alpha}\ln y_{i:m:n}+R_J^*T^{\alpha}\ln T$, for Case-I and Case-II respectively. Note that

$$\lambda^{\alpha}=\frac{W(\alpha)}{d}=u(\alpha)\ \ (say) \tag{12}$$

and the MLE of α can be obtained by solving

$$\alpha=h(\alpha), \tag{13}$$

where

$$h(\alpha)=\frac{d}{-\sum_{i=1}^{d}\ln y_{i:m:n}+\frac{1}{u(\alpha)}W(\alpha)}.$$

A simple iterative scheme is proposed to obtain the MLE of α from (13). Starting with an initial guess of α, for example, $\alpha^{(0)}$, obtain $\alpha^{(1)}=h(\alpha^{(0)})$ and proceed in this way to obtain $\alpha^{(n+1)}=h(\alpha^{(n)})$. The iterative procedures stops when $\left|\alpha^{(n+1)}-\alpha^{(n)}\right|<\in$, which is some pre-assigned tolerance limit. Once the MLE of α is obtained the MLE of λ can be obtained from (12). Since the MLE's, when they exist, are not in compact forms, the following approximate MLE's and its' explicit expressions are proposed.

Approximate Maximum Likelihood Estimators (AMLEs)

Using the following notations $x_{i:m:n}=\ln y_{i:m:n}$ and $S=\ln T$, the likelihood equation of the observed data $x_{i:m:n}$ for Case-I is

$$l(\mu,\sigma)=\frac{1}{\sigma^m}\prod_{i=1}^{m}\left[n-\sum_{k=1}^{i-1}(1+R_K)\right]g(z_{i:m:n})\left(\bar{G}(z_{i:m:n})\right)^{R_i}, \tag{14}$$

and for Case II is

$$l(\mu,\sigma)=\frac{1}{\sigma^J}\prod_{i=1}^{J}\left[n-\sum_{k=1}^{i-1}(1+R_K)\right]g(z_{i:m:n})\left(\bar{G}(z_{i:m:n})\right)^{R_i}\left(\bar{G}(V)\right)^{R_J^*} \tag{15}$$

where ,

$z_{i:m:n}=\left(x_{i:m:n}-\mu\right)/\sigma$, $V=\left(S-\mu\right)/\sigma$, $g(x)=e^{x-e^x}$, $\bar{G}(x)=e^{-e^x}$, $\mu=\ln\lambda$ and $\sigma=1/\alpha$.

Ignoring the constant term, the following log-likelihood results from (15) is

$$L(\mu,\sigma)=\ln\left[l(\mu,\sigma)\right]=-m\ln\sigma+\sum_{i=1}^{m}\ln\left(g(z_{i:m:n})\right)+\sum_{i=1}^{m}R_i\ln\left(\bar{G}(z_{i:m:n})\right). \tag{16}$$

From (16) the following approximate MLE's of μ and σ are obtained (see Appendix 1),

$$\tilde{\mu}=\frac{\left(c_1-c_2-m\right)\tilde{\sigma}+d_1}{c_1},\ \tilde{\sigma}=\frac{-B+\sqrt{B^2-4AC}}{2A} \tag{17}$$

where

$c_1=\sum_{i=1}^{m}D_ie^{\mu_i}$, $c_2=\sum_{i=1}^{m}D_i\mu_ie^{\mu_i}$, $d_1=\sum_{i=1}^{m}D_iX_{i:m:n}e^{\mu_i}$, $d_2=\sum_{i=1}^{m}D_iX_{i:m:n}^2e^{\mu_i}$, $d_3=\sum_{i=1}^{m}D_i\mu_iX_{i:m:n}e^{\mu_i}$, $A=m\,c_1$, $B=c_1\left(d_3+m\bar{X}\right)-d_1\left(c_2+m\right)$, $C=d_1^2-c_1d_2$, $\mu_i=G^{-1}(p_i)=\ln(-\ln q_i)$, $p_i=i/(n+1)$, $q_i=1-p_i$ and $D_i=1+R_i$, for $i=1,\cdots,m$.

For Case-II, ignoring the constant term, the log-likelihood is obtained as

$$L(\mu,\sigma)=\ln\left[l(\mu,\sigma)\right]=-J\ln\sigma+\sum_{i=1}^{J}\ln\left(g(z_{i:m:n})\right)+\sum_{i=1}^{J}R_i\ln\left(\bar{G}(z_{i:m:n})\right)+R_J^*\ln\bar{G}(V). \tag{18}$$

In this case the approximate MLE's are (see Appendix 2)

$$\tilde{\mu}=\frac{\left(c_1'-c_2'-J\right)\tilde{\sigma}+d_1'}{c_1'},\quad \tilde{\sigma}=\frac{-B'+\sqrt{B'^2-4A'C'}}{2A'} \tag{19}$$

where,

$c_1'=\sum_{i=1}^{J}D_ie^{\mu_i}+R_J^*e^{\mu_J^*}$,

$c_2'=\sum_{i=1}^{J}D_i\mu_ie^{\mu_i}+R_J^*\mu_J^*e^{\mu_J^*}$,

$d_1'=\sum_{i=1}^{J}D_iX_{i:m:n}e^{\mu_i}+R_J^*Se^{\mu_J^*}$,

$d_2'=\sum_{i=1}^{J}D_iX_{i:m:n}^2e^{\mu_i}+R_J^*S^2e^{\mu_J^*}$,

$d_3'=\sum_{i=1}^{J}D_i\mu_iX_{i:m:n}e^{\mu_i}+R_J^*\mu_J^*Se^{\mu_J^*}$,

$A'=Jc_1'$, $B'=c_1'\left(d_3'+J\bar{X}\right)-d_1'\left(c_2'+J\right)$,

$C'=d_1'^2-c_1'd_2'$.

Here μ_i and D_i are the same as above for $i=1,..,J$, $\mu_J^*=G^{-1}(p_J^*)=\ln(-\ln q_J^*)$, $p_J^*=(p_J+p_{J+1})/2$, and $q_J^*=1-p_J^*$.

Results

Because the performance of the different methods cannot be compared theoretically, Monte Carlo simulations are used to compare the performances of the different methods proposed for different parameter values and for different sampling schemes. The term different sampling schemes mean different sets of $R_i's$ and different T values. The performances of the MLEs and AMLEs estimators of the unknown parameters are compared in terms of their biases and mean squared errors (MSEs) for different censoring schemes. The average lengths of the asymptotic confidence intervals and their coverage percentages are also compared. All computations were performed using a Pentium IV processor and a FORTRAN-77 program. In all cases the random deviate generator RAN2 was used as proposed in Press, et al. (1991).

Because λ is the scale parameter, all cases $\lambda=1$ have been taken in without loss of generality. For simulation purposes, the results are presented when T is of the form $T^{1/\alpha}$. The reason for choosing T in that form is as follows: if $\hat{\alpha}$ represents the MLE or AMLE of α, then the distribution of $\hat{\alpha}/\alpha$ becomes independent of α in the case for $\lambda=1$. For that purpose the result is reported only for $\alpha=1$ without loss of generality, however, these results can be used for any other α also.

Type-II progressively hybrid censored data is generated for a given set $n, m, R_1, ..., R_m$ and T by using the following transformation for exponential distribution as suggested in Balakrishnan and Aggarwala (2000).

$$\begin{aligned} Z_1 &= nE_{1:m:n} \\ Z_2 &= (n - R_1 - 1)(E_{2:m:n} - E_{1:m:n}) \\ &\vdots \\ Z_m &= (n - R_1 - ... - R_{m-1} - m + 1)(E_{m:m:n} - E_{m-1:m:n}) \end{aligned} \quad (20)$$

It is known that if $E_i's$ are i.i.d. standard exponential, then the spacing $Z_i's$ are also i.i.d. standard exponential random variables. From (20) it follows that

$$\begin{aligned} E_{1:m:n} &= \frac{1}{n} Z_1 \\ E_{2:m:n} &= \frac{1}{n - R_1 - 1} Z_2 + \frac{1}{n} Z_1 \\ &\vdots \\ E_{m:m:n} &= \frac{1}{n - R_1 - ... - R_{m-1} - m + 1} Z_m + ... + \frac{1}{n} Z_1. \end{aligned} \quad (21)$$

Using (21) and parameters α and λ, Type-II progressively hybrid censored data for the Weibull distribution can be generated for a given $n, m, R_1, .., R_m$, $Y_{1:m:n}, ..., Y_{m:m:n}$. If $Y_{m:m:n} < T$, then Case I results and the corresponding sample is $\{(Y_{1:m:n}, R_1), ..., (Y_{m:m:n}, R_m)\}$. If $Y_{m:m:n} > T$, then Case II results and J is found such that $Y_{J:m:n} < T < Y_{J+1:m:n}$. The corresponding Type-II hybrid censored sample is $\{(Y_{1:m:n}, R_1), ..., (Y_{J:m:n}, R_J)\}$ and R_J^*, where R_J^* is same as defined before.

Consider different n, m and T. Two different sampling schemes have been used, namely,

Scheme 1:

$$R_1 = ... = R_{m-1} = 0 \text{ and } R_m = n - m.$$

Scheme 2:

$$R_1 = ... = R_{m-1} = 1 \text{ and } R_m = n - 2m + 1.$$

Note that Scheme 1 is the conventional Type-II censoring scheme and Scheme 2, is a typical progressive censoring scheme. In each case the MLEs and AMLEs are computed as estimates of the unknown parameters. The 95% asymptotic confidence intervals are calculated based on MLEs by replacing the MLEs by AMLEs. The process was replicated 1,000 times. Average estimates, MSEs and average confidence lengths with coverage percentages were reported in Tables 1-8.

Based on Tables 1-4 (for MLEs) and Tables 5-8 (for AMLEs), the following observations are made: As expected, for fixed n, as m increases the biases and the MSEs decrease for both α and λ, however, for fixed m as n increases this may not be true. This shows that the effective sample size (m) plays an important role when considering the actual sample size (n). It is also observed that the MLEs for schemes 1 and 2 behave quite similarly in terms of biases and MSEs, unless both n and m are small. The performances in terms of biases and MSEs improve as T increases. Similar results are also observed for AMLEs.

Comparing different confidence intervals in terms of average lengths and coverage probabilities, it is generally observed that both the methods work well even for small n and m. For both methods, it is observed that the average confidence lengths decrease as n increases for fixed m, or vice versa. For both the MLE and AMLE methods, scheme 1 and scheme 2 behave very similarly although the confidence intervals for scheme 1 tend to be slightly shorter than scheme 2.

Data Analysis

Kundu and Joarder (2006) analyzed the following two data sets obtained from Lawless (1982) using exponential distributions.

Data Set 1

In this case $n = 36$ and, if $m = 10, T = 2600, R_1 = R_2 = ... = R_9 = 2, R_{10} = 8,$ then the Type II progressively hybrid censored

sample is: 11, 35, 49, 170, 329, 958, 1,925, 2,223, 2,400, 2,568. From the above sample data, $D = m = 10$ is obtained, which yields α and λ of based on MLEs and AMLEs are $(\hat{\alpha}=6.29773\times10^{-1}, \hat{\lambda}=8113.80176)$, $(\tilde{\alpha}=6.33116\times10^{-1}, \tilde{\lambda}=6511.83036)$ respectively. Using the above estimates the 95% asymptotic confidence interval for α and λ is obtained based on MLEs and AMLEs which are

$$(6.29773\times10^{-1}, 6.29882\times10^{-1}),$$
$$(8113.40869, 8114.19482)$$

and

$$(6.33116\times10^{-1}, 6.33176\times10^{-1}),$$
$$(6511.4344, 6512.2264)$$

respectively.

Data Set 2

Consider $m = 10,\ T = 2000,$ and R_i's are same as Data Set 1. In this case the progressively hybrid censored sample obtained as: 11, 35, 49, 170, 329, 958, 1,925 and $D = J = 7.$ The MLE and AMLEs of α and λ are $(\hat{\alpha} = 4.77441\times10^{-1},\ \hat{\lambda} = 25148.8613)$ and $(\tilde{\alpha} = 4.77589\times10^{-1},\ \tilde{\lambda} = 23092.3759)$ respectively. From the above estimates the 95% asymptotic confidence intervals are obtained for α and λ based on MLEs and AMLEs, which are

$$(4.77383\times10^{-1}, 4.77499\times10^{-1}),$$
$$(25148.5078, 25149.2148)$$

and

$$(4.77529\times10^{-1}, 4.77649\times10^{-1}),$$
$$(23092.0219, 23092, 7299)$$

respectively.

In both cases it is clear that if the tested hypothesis is $H_0 : \alpha = 1$, it will be rejected, this implies that in this case the Weibull distribution should be used rather than exponential.

Conclusion

This article discussed the Type-II progressively hybrid censored data for the two parameters Weibull distribution. It was observed that the maximum likelihood estimator of the shape parameter could be obtained by using an iterative procedure. The proposed approximate maximum likelihood estimators of the shape and scale parameters could be obtained in explicit forms. Although the exact confidence intervals could not be constructed, it was observed that the asymptotic confidence intervals work reasonably well for MLEs. Although the frequentest approach was used, Bayes estimates and credible intervals can also be obtained under suitable priors along the same line as Kundu (2007).

References

Arnold, B. C., & Balakrishnan, N. (1989). *Relations bounds and approximations for order statistics, lecture notes in statistics (53)*. New York: Springer Verlag.

Balakrishnan, N. (2007). Progressive censoring: An apprisal (with discussions). In press. To appear in *TEST*.

Balakrishnan, N., & Aggrwala, R. (2000). *Progressive censoring: Theory, methods and applications*. Boston, MA: Birkhauser.

Balakrishnan, N., & Varadan, J. (1991). Approximate MLEs for the location and scale parameters of the extreme value distribution with censoring. *IEEE Transactions on Reliability*, *40*, 146-151.

Childs, A., Chandrasekhar, B., & Balakrishnan, N. (2007). *Exact likelihood inference for an exponential parameter under progressively hybrid schemes: Statistical models and methods for biomedical and technical systems*. F. Vonta, M. Nikulin, N. Limnios, & C. Huber-Carod (Eds.). Boston, MA: Birkhauser.

Cohen, A. C. (1963). Progressively censored samples in life testing. *Technometrics*, *5*, 327-329.

Cohen, A. C. (1966). Life-testing and early failure. *Technometrics*, *8*, 539-549.

David, H. A. (1981). *Order Statistics* (*2nd Ed.*). New York: John Wiley and Sons.

Kundu, D., & Joarder, A. (2006). Analysis of type II progressively hybrid censored data. *Computational Statistics and Data Analysis*, *50*, 2509-2528.

Kundu, D. (2007). On hybrid censored Weibull distribution. *Journal of Statistical Planning and Inference*, *137*, 2127-2142.

Table 1: MLE Estimate for T = 0.75

N.M.		Scheme 1	Scheme 2
30, 15	α	1.0968(0.0862), 1.2913(94.5)	1.0751(0.0838), 1.1898(93.5)
	λ	1.0358(0.1611),1.6019(89.1)	1.0760(0.2937), 1.6015(88.8)
40, 20	α	1.0898(0.0623), 1.0099(96.6)	1.0750(0.0626), 1.0167(94.9)
	λ	1.0111(0.0934), 1.2453(92.3)	1.413(0.1321), 1.3662(90.7)
60, 20	α	1.1046(0.0644), 1.1701(92.3)	1.0916(0.0554), 1.0255(94.7)
	λ	0.9777(0.0962), 1.6693(88.5)	0.9842(0.0902), 1.4432(91.2)
60, 30	α	1.0473(0.0342), 0.7386(96.5)	1.0385(0.0364), 0.7681(95.1)
	λ	1.0109(0.0653), 0.9055(92.7)	1.0350(0.0962), 1.0315(90.9)
80, 30	α	1.0566(0.0344), 0.7918(95.6)	1.0435(0.0302), 0.7074(96.1)
	λ	0.9913(0.0633), 1.0782(92.5)	1.0081(0.0731), 0.9630(92.6)
80, 40	α	1.0401(0.0252), 0.6275(97.3)	1.0301(0.0269), 0.6501(95.6)
	λ	1.0060(0.0449), 0.7670(93.2)	1.0261(0.0614), 0.8732(91.7)
100, 40	α	1.0471(0.0256), 0.6620(97.4)	1.0323(0.0219), 0.5932(96.4)
	λ	0.9904(0.0406), 0.878(93.4)	1.0096(0.0465), 07985(93.8)
100, 50	α	1.0369(0.0209), 05544(96.2)	1,0281(0.0232), 0.5811(95.6)
	λ	0.9996(0.0292), 0.6760(93.6)	1.0185(0.0418), 0.7800(93.0)

Table 2: MLE Estimate for T = 1.00

N.M.		Scheme 1	Scheme 2
30, 15	α	1.1102(0.0841), 1.2367(96.0)	1.0719(0.0730), 1.0287(95.6)
	λ	0.9982(0.1171), 1.5080(92.1)	1.0383(0.1397), 1.3333(91.2)
40, 20	α	1.0983(0.0600), 0.9891(97.7)	1.0704(0.0518), 0.8833(96.4)
	λ	0.9864(0.0629), 1.2035(93.7)	1.0179(0.0817), 1.1445(92.1)
60, 20	α	1.1046(0.0644), 1.1701(92.3)	1.0933(0.0550), 1.0249(95.2)
	λ	0.9781(0.0982), 1.6692(88,5)	0.9776(0.0793), 1.4394(91.6)
60, 30	α	1.0539(0.0329), 0.7320(97.0)	1.0358(0.0291), 0.6855(95.9)
	λ	0.9945(0.0510), 0.8876(94.2)	1.0157(0.0616), 0.8892(92.3)
80, 30	α	1.0567(0.0344), 0.7918(95.7)	1.0487(0.0291), 0.7049(96.9)
	λ	0.9906(0.0605), 1.0781(92.5)	0.9926(0.0553), 0.9508(93.9)
80, 40	α	1.0456(0.0246), 06214(97.8)	0.0313(0.0225), 0.5879(97.0)
	λ	0.9927(0.0331), 0.7531(94.1)	1.0110(0.0429), 0.7624(92.2)
100, 40	α	1.0473(0.0255), 0.6621(97.4)	1.0396(0.0211), 0.5788(97.4)
	λ	0.9895(0.0385), 0.8781(93.4)	0.9936(0.0364), 0.7655(94.0)
100, 50	α	1.0397(0.0205), 0.5493(96.9)	1.0252(0.0190), 0.5216(94.7)
	λ	0.9927(0.0243), 0.6653(94.0)	1.0120(0.0301), 0.6773(93.5)

Table 3: MLE Estimate for T = 1.50

N.M.		Scheme 1	Scheme 2
30, 15	α	1.1130(0.0833),1.2367(96.3)	1.0727(0.0630), 0.9343(95.8)
	λ	0.9857(0.0820), 1.5075(92.7)	1.0196(0.1079), 1.2004(92.7)
40, 20	α	1.0992(0.0599), 0.9886(97.8)	1.0682(0.0430), 0.7962(97.5)
	λ	0.9841(0.0600), 1.2025(93.6)	1.0025(0.0593), 1.0237(94.4)
60, 20	α	1.1046(0.0644), 1.1701(92.3)	1.0932(0.0550), 1.0248(95.2)
	λ	0.9781(0.0982), 1.6692(88.5)	0.9779(0.0807), 1.4394(91.6)
60, 30	α	1.0544(0.0327), 0.7320(97.2)	1.0366(0.0259), 0.6251(94.9)
	λ	0.9920(0.0451), 0.8875(94.2)	1.0054(0.0498), 0.8042(93.0)
80, 30	α	1.0567(0.0344), 0.7918(95.7)	1.0492(0.0288), 0.7051(97.0)
	λ	0.9906(0.0605), 1.0781(92.5)	0.9900(0.0503), 09508(93.8)
80, 40	α	1.0458(0.0245), 0.6215(97.8)	1.0308(0.0192), 0.5357(96.8)
	λ	0.9919(0.0312),0.7531) (94.1)	1.0031(0.0319), 0.6896(93.6)
100, 40	α	1.0473(0.0255),0.6621(97.4)	1.0407(0.0209), 0.5785(97.7)
	λ	0.9895(0.0385), 08781(93.4)	0.9901(0.0322), 0.7645(94.0)
100, 50	α	1.0397(0.0205), 0.5492(96.9)	1.0277(0.0156), 0.4768(94.7)
	λ	0.9928(0.0243), 0.6652(94.0)	1.0008(0.0231), 0.6138(94.7)

Table 4: MLE Estimate for T = 2.00

N.M.		Scheme 1	Scheme 2
30, 15	α	1.1130(0.0833),1.2367(96.3)	1.0754(0.062), 0.9106(96.1)
	λ	0.9857(0.0820), 1.5075(92.7)	1.0045(0.0882), 1.1750(92.6)
40, 20	α	1.0992(0.0599), 0.9886(97.8)	1.0695(0.0423),0.7720(95.8)
	λ	0.9841(0.0600), 1.2025(93.6)	0.9966(0.0538), 0.9983(94.6)
60, 20	α	1.1046(0.0644), 1.1701(92.3)	1.0932(0.0550),1.0248(95.2)
	λ	0.9781(0.0982), 1.6692(88.5)	0.9779(0.0807), 1.4394(91.6)
60, 30	α	1.0544(0.0327), 0.7320(97.2)	1.0379(0.0248), 0.6054(95.6)
	λ	0.9920(0.0451), 0.8875(94.2)	1.0004(0.0433), 0.7836(94.2)
80, 30	α	1.0567(0.0344), 0.7918(95.7)	1.0492(0.0288), 0.7051(97.0)
	λ	0.9906(0.0605), 1.0781(92.5)	0.9900(0.0503), 0.9508(93.8)
80, 40	α	1.0458(0.0245), 0.6215(97.8)	1.0321(0.0176),0.5179(96.6)
	λ	0.9919(0.0312),0.7531) (94.1)	0.9986(0.0283), 0.6709(94.1)
100, 40	α	1.0473(0.0255),0.6621(97.4)	1.0407(0.0209), 0.5785(97.7)
	λ	0.9895(0.0385), 08781(93.4)	0.9901(0.0322), 0.7645(94.0)
100, 50	α	1.0397(0.0205), 0.5492(96.9)	1.0286(0.0149), 0.4608(94.6)
	λ	0.9928(0.0243), 0.6652(94.0)	0.9986(0.0219), 0.5969(94.1)

Table 5: Approximate MLE Estimate for T = 0.75

N.M.		Scheme 1	Scheme 2
30, 15	α	1.0873(0.0847), 1.2073(94.2)	1.0814(0.0889), 1.2070(93.3)
	λ	1.0354(0.1640), 1.4941(89.1)	1.0104(0.3033), 1.5970(88.6)
40, 20	α	1.0832(0.0615), 0.9924(96.2)	1.0837(0.06559), 1.0651(95.4)
	λ	1.01103(0.0941), 1.2224(92.2)	0.9752(0.1333), 1.4107(91.3)
60, 20	α	1.0998(0.0638), 1.1226(92.1)	1.0915(0.0554), 1.0236(94.5)
	λ	0.9792(0.0968), 1.6005(88.4)	0.9435(0.0823), 1.4270(92.8)
60, 30	α	1.0432(0.0340), 0.7349(96.5)	1.0486(0.0386), 0.7959(95.5)
	λ	1.0102(0.0655), 0.900(92.7)	0.9679(1.0962), 1.0530(92.1)
80, 30	α	1.0533(0.0342), 07870(95.5)	1.0492(0.0308), 0.7288(96.4)
	λ	0.9920(0.0635), 1.0714(92.5)	0.9520(0.0688), 0.9797(94.7)
80, 40	α	1.0372(0.0251), 0.6253(97.1)	1.0409(0.0284), 0.6735(96.3)
	λ	1,0054(0.0450), 0.7640(93.2)	0.9588(0.0620), 0.8914(92.2)
100, 40	α	1,0447(0,0255), 0,6593(97.2)	1.0417(0.0227), 0.6140(97.4)
	λ	0.9906(0.0407), 08743(93.4)	0.9463(0.0453), 08151(94.8)
100, 50	α	1.0346(0.0208), 0.5529(96.2)	1.0392(0.0243), 0.6029(96.4)
	λ	0.9991(0.0292), 0.6739(93.6)	0.9512(0.0421), 0.7976(93.9)

Table 6: Approximate MLE Estimate for T = 1.00

N.M.		Scheme 1	Scheme 2
30, 15	α	1.1003(0.827),1.1683(95.3)	1.0921(0.0811), 1.0824(96.1)
	λ	0.9968(0.1177), 1.4208(92.0)	0.9378(0.1395), 1.3736(92.3)
40, 20	α	1.0916(0.0592), 0.9731(97.4)	1.0936(0.0582), 0.9316(97.0)
	λ	0.9851(0.0628), 1.1827(93.8)	0.9175(0.0809), 1.1.822(94.3)
60, 20	α	1.0998(0.0638), 1.1226(92.1)	1.0933(0.0550),1.0232(94.9)
	λ	0.9797(0.0989), 1.6004(88.4)	0.9359(0.0703), 1.4234(93.4)
60, 30	α	1.0497(0.0327), 0.7283(97.0)	1.0586(0.0326), 0.7217(96.7)
	λ	0.9936(0.0510), 0.8827(94.2)	0.9150(0.0608), 0.9169(92.7)
80, 30	α	1.0534(0.0342), 0.7870(95.6)	1.0555(0.0296), 0.7275(97.0)
	λ	0.9912(0.0607), 1.0712(92.5)	0.9309(0.0476), 0.9685(96.6)
80, 40	α	1.0426(0.0244), 0.6193(97.7)	1.0546(0.0251), 0.6189(97.2)
	λ	0.9921(0.0330), 0.7502(94.2)	0.9102(0.0429), 0.7863(92.3)
100, 40	α	1.0448(0.0254), 0.6594(97.2)	1.0518(0.0218), 0.6009(98.0)
	λ	0.9897(0.0385), 0.8743(93.4)	0.9183(0.0311), 0.7825(95.7)
100, 50	α	1.0374(0.0204), 0.5478(96.8)	1.0484(0.0211), 0.5489(95.6)
	λ	0.9922(0.0243), 0.6633(94.0)	0.9112(0.0299), 0.6984(93.7)

Table 7: Approximate MLE Estimate for T = 1.50

N.M.		Scheme 1	Scheme 2
30, 15	α	1.1030(0.0819), 1.1681(95.7)	1.1153(0.0735), 1.0048(96.1)
	λ	0.9841(0.0820), 1.4202(92.7)	0.8709(0.0978), 1.2553(93.6)
40, 20	α	1.0925(0.0592), 0.9726(97.6)	1.1158(0.0519), 0.8603(96.9)
	λ	0.9827(0.0598), 1.1818(93.7)	0.8541(0.0524), 1.0754(95.5)
60, 20	α	1.0998(0.0638), 1.1226(92.1)	1.0932(0.0550), 1.0231(94.9)
	λ	0.9797(0.0989), 1.6004(88.4)	0.9361(0.0712), 1.4233(93.4)
60, 30	α	1.0502(0.0325), 0.7284(97.1)	1.0832(0.0313), 0.6753(95.1)
	λ	0.9910(0.0450), 0.8826(94.2)	0.8563(0.0443), 0.8445(92.8)
80, 30	α	1.0534(0.0342), 0.7870(95.6)	1.0560(0.0293), 0.7277(97.2)
	λ	0.9912(0.0607), 1.0712(92.5)	0.9280(0.0424), 0.9684(96.6)
80, 40	α	1.0428(0.0244), 0.6193(97.7)	1.0778(0.0232), 0.5787(95.8)
	λ	0.9912(0.0311), 0.7502(94.2)	0.8540(0.0287), 0.7242(92.0)
100, 40	α	1.0448(0.0254), 0.6594(97.2)	1.0532(0.0215), 0.6009(98.3)
	λ	0.9897(0.0385), 0.8743(93.4)	0.9136(0.0262), 0.7816(96.8)
100, 50	α	1.0374(0.0204), 0.5477(96.8)	1.0748(0.0188), 0.5152(94.4)
	λ	0.9922(0.0243), 0.6632(94.0)	0.8514(0.0205), 0.6447(91.6)

Table 8: Approximate MLE Estimate for T = 2.00

N.M.		Scheme 1	Scheme 2
30, 15	α	1.1030(0.0819), 1.1681(95.7)	1.1337(0.0710), 0.9924(96.1)
	λ	0.9841(0.0820), 1.4202(92.7)	0.8327(0.0690), 1.2439(95.0)
40, 20	α	1.0925(0.0592), 0.9726(97.6)	1.1326(0.0512), 0.8454(96.0)
	λ	0.9827(0.0598), 1.1818(93.7)	0.8245(0.0414), 1.0610(96.1)
60, 20	α	1.0998(0.0638), 1.1226(92.1)	1.0932(0.0550), 1.0231(94.9)
	λ	0.9797(0.0989), 1.6004(88.4)	0.9361(0.0712), 1.4233(93.4)
60, 30	α	1.0502(0.0325), 0.7284(97.1)	1.0990(0.0302), 0.6630(94.8)
	λ	0.9910(0.0450), 0.8826(94.2)	0.8269(0.0336), 0.8319(92.5)
80, 30	α	1.0534(0.0342), 0.7870(95.6)	1.0560(0.0293), 0.7277(97.2)
	λ	0.9912(0.0607), 1.0712(92.5)	0.9280(0.0424), 0.9684(96.6)
80, 40	α	1.0428(0.0244), 0.6193(97.7)	1.0946(0.0217), 0.5678(94.6)
	λ	0.9912(0.0311), 0.7502(94.2)	0.8247(0.0222), 0.7129(90.8)
100, 40	α	1.0448(0.0254), 0.6594(97.2)	1.0532(0.0215), 0.6009(98.3)
	λ	0.9897(0.0385), 0.8743(93.4)	0.9136(0.0262), 0.7816(96.8)
100, 50	α	1.0374(0.0204), 0.5477(96.8)	1.0916(0.0185), 0.5053(92.4)
	λ	0.9922(0.0243), 0.6632(94.0)	0.8245(0.0170), 0.6346(89.4)

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. New York: Wiley.

Mann, N. R. (1971). Best linear invariant estimation for Weibull parameters under progressive censoring. *Technometrics*, *13*, 521-533.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1991). *Numerical recipes: The art of scientific computing*. Cambridge, U.K.: Cambridge University Press.

Thomas, D. R., & Wilson, W. M. (1972). Linear order statistics estimation for the two-parameter Weibull and extreme value distribution from Type-II progressively censored samples. *Technometrics*, *14*, 679-691.

Appendix 1

For case-I, taking derivatives with respect to μ and σ of $L(\mu,\sigma)$ as defined in (16), results in

$$\frac{\partial L(\mu,\sigma)}{\partial \mu}=\frac{1}{\sigma}\left[\sum_{i=1}^{m} R_i \frac{g(z_{i:m:n})}{\bar{G}(z_{i:m:n})}-\sum_{i=1}^{m}\frac{g'(z_{i:m:n})}{g(z_{i:m:n})}\right]=0 \tag{22}$$

$$\frac{\partial L(\mu,\sigma)}{\partial \mu}=\frac{1}{\sigma}\left[\sum_{i=1}^{m} R_i z_{i:m:n} \frac{g(z_{i:m:n})}{\bar{G}(z_{i:m:n})}-\sum_{i=1}^{m} z_{i:m:n}\frac{g'(z_{i:m:n})}{g(z_{i:m:n})}-m\right]=0. \tag{23}$$

Clearly, (22) and (23) do not have explicit analytical solutions. Consider a first-order Taylor approximation to $g'(z_{i:m:n})/g(z_{i:m:n})$ and $g(z_{i:m:n})/\bar{G}(z_{i:m:n})$ by expanding around the actual mean μ_i of the standardized order statistic $Z_{i:m:n}$, where $\mu_i = G^{-1}(p_i) = \ln(-\ln q_i)$, and $p_i = i/(n+1)$, $q_i = 1-p_i$ for $i=1,...,m$, similar to Balakrishnan and Varadan (1991), David (1981) or Arnold and Balakrishnan (1989). Otherwise, the necessary procedures for obtaining μ_i, $i=1,..,m$, were made available by Mann (1971) and Thomas and Wilson (1972). Note that for $i=1,...,m$

$$g'(z_{i:m:n})/g(z_{i:m:n}) \approx \alpha_i - \beta_i z_{i:m:n} \tag{24}$$

$$g(z_{i:m:n})/\bar{G}(z_{i:m:n}) \approx 1-\alpha_i + \beta_i z_{i:m:n} \tag{25}$$

where,

$$\alpha_i = \frac{g'(\mu_i)}{g(\mu_i)} - \mu_i\left[\frac{g''(\mu_i)}{g'(\mu_i)} - \left(\frac{g'(\mu_i)}{g'(\mu_i)}\right)^2\right]$$
$$= 1+\ln q_i(1-\ln(-\ln q_i)),$$
$$\beta_i = \left[-\frac{g''(\mu_i)}{g'(\mu_i)}+\left(\frac{g'(\mu_i)}{g(\mu_i)}\right)^2\right] = -\ln q_i$$

Using the approximation (24) and (25) in (22) and (23), results get

$$\left[\sum_{i=1}^{m} D_i e^{\mu_i} - \sum_{i=1}^{m} D_i \mu_i e^{\mu_i} - m\right]\sigma + \sum_{i=1}^{m} D_i X_{i:m:n} e^{\mu_i} - \mu \sum_{i=1}^{m} D_i e^{\mu_i} = 0 \tag{26}$$

and

$$\left[m\sum_{i=1}^{m} D_i e^{\mu_i}\right]\sigma^2 +$$
$$\left[\sum_{i=1}^{m} D_i e^{\mu_i}\left(\sum_{i=1}^{m} D_i \mu_i X_{i:m:n} e^{\mu_i} + m\bar{X}\right)\right]\sigma +$$
$$\left[\sum_{i=1}^{m} D_i X_{i:m:n} e^{\mu_i}\right]^2$$
$$-\left[\sum_{i=1}^{m} D_i X_{i:m:n} e^{\mu_i}\left(\sum_{i=1}^{m} D_i \mu_i e^{\mu_i} + m\right)\right]\sigma$$
$$-\left[\sum_{i=1}^{m} D_i e^{\mu_i}\right]\left[\sum_{i=1}^{m} D_i X_{i:m:n}^2 e^{\mu_i}\right] = 0 \tag{27}$$

The above two equations (26) and (27) can be written as

$$(c_1 - c_2 - m)\sigma + d_1 - \mu c_1 = 0 \tag{28}$$

$$A\sigma^2 + B\sigma + C = 0 \tag{29}$$

where

$c_1 = \sum_{i=1}^{m} D_i e^{\mu_i}$, $c_2 = \sum_{i=1}^{m} D_i \mu_i e^{\mu_i}$, $d_1 = \sum_{i=1}^{m} D_i X_{i:m:n} e^{\mu_i}$,

$$d_2 = \sum_{i=1}^{m} D_i X_{i:m:n}^2 e^{\mu_i},$$

$d_3 = \sum_{i=1}^{m} D_i \mu_i X_{i:m:n} e^{\mu_i}$, $A = mc_1$, $B = c_1(d_3 + m\bar{X}) - d_1(c_2 + m)$,

$$C = d_1^2 - c_1 d_2$$

and $D_i = 1 + R_i$ for $i = 1,..,m$. The solution to the preceding equations yields the approximate MLE's are

$$\tilde{\mu} = \frac{(c_1 - c_2 - m)\tilde{\sigma} + d_1}{c_1} \tag{30}$$

$$\tilde{\sigma} = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{31}$$

Consider only positive root of σ; these approximate estimators are equivalent but not unbiased. Unfortunately, it is not possible to compute the exact bias of $\tilde{\mu}$ and $\tilde{\sigma}$ theoretically because of intractability encountered in finding the expectation of $\sqrt{B^2 - 4AC}$.

Appendix 2

For case-II, taking derivatives with respect to μ and σ of $L(\mu, \sigma)$ as defined in (18), gives (similar to Case-I)

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma}\left[\sum_{i=1}^{J} R_i \frac{g(z_{i:m:n})}{\bar{G}(z_{i:m:n})} - \sum_{i=1}^{J} \frac{g'(z_{i:m:n})}{g(z_{i:m:n})} + R_J^* \frac{g(V)}{\bar{G}(V)}\right] = 0 \tag{32}$$

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma}\left[\sum_{i=1}^{J} R_i z_{i:m:n} \frac{g(z_{i:m:n})}{\bar{G}(z_{i:m:n})} - \sum_{i=1}^{J} z_{i:m:n} \frac{g'(z_{i:m:n})}{g(z_{i:m:n})} + R_J^* V \frac{g(V)}{\bar{G}(V)} - J\right] = 0. \tag{33}$$

Here again consider the first-order Taylor approximation to $g'(z_{i:m:n})/g(z_{i:m:n})$ and $g(z_{i:m:n})/\bar{G}(z_{i:m:n})$ by expanding around the actual mean μ_i of the standardized order statistic $Z_{i:m:n}$, where μ_i's are defined in Appendix 1. Here $g(V)/\bar{G}(V)$ is also exploded in the Taylor series around the point μ_J^*, where

$$\mu_J^* = G^{-1}\left(p_J^*\right) = \ln\left(-\ln q_J^*\right), \quad p_J^* = \left(p_J + p_{J+1}\right)/2$$

and $q_J^* = 1 - p_J^*$.

Note that

$$\frac{g'(V)}{g(V)} \approx \alpha_J^* - \beta_J^* V \tag{34}$$

$$\frac{g(V)}{\bar{G}(V)} \approx 1 - \alpha_J^* + \beta_J^* V \tag{35}$$

where

$$\alpha_J^* = \frac{g'(\mu_J^*)}{g(\mu_J^*)} - \mu_J^*\left[\frac{g''(\mu_J^*)}{g'(\mu_J^*)} - \left(\frac{g'(\mu_J^*)}{g(\mu_J^*)}\right)^2\right] = 1 + \ln q_J^*\left(1 - \ln\left(-\ln q_J^*\right)\right),$$

$$\beta_J^* = \left[-\frac{g''(\mu_J^*)}{g'(\mu_J^*)} + \left(\frac{g'(\mu_J^*)}{g(\mu_J^*)}\right)^2\right] = -\ln q_J^*$$

Using the approximation (24), (25), (34) and (35) in (32) and (33) gives

$$\left[\left(\sum_{i=1}^{J} D_i e^{\mu_i} + R_J^* e^{\mu_J^*}\right) - \left(\sum_{i=1}^{J} D_i \mu_i e^{\mu_i} + R_J^* \mu_J^* e^{\mu_J^*}\right) - J\right]\sigma$$
$$\left[\sum_{i=1}^{J} D_i X_{i:m:n} e^{\mu_i} + R_J^* S e^{\mu_J^*}\right] - \mu\left[\sum_{i=1}^{J} D_i e^{\mu_i} + R_J^* e^{\mu_J^*}\right] = 0 \tag{36}$$

and

$$\left[J\sum_{i=1}^{J} D_i e^{\mu_i} + R_J^* e^{\mu_J^*}\right]\sigma^2 + \left[\left(\sum_{i=1}^{J} D_i e^{\mu_i} + R_J^* e^{\mu_J^*}\right)\left(\sum_{i=1}^{J} D_i \mu_i X_{i:m:n} e^{\mu_i} + R_J^* \mu_J^* S e^{\mu_J^*} + J\bar{X}\right)\right]\sigma$$
$$-\left[\left(\sum_{i=1}^{J} D_i X_{i:m:n} e^{\mu_i} + R_J^* S e^{\mu_J^*}\right)\left(\sum_{i=1}^{J} D_i \mu_i e^{\mu_i} + R_J^* \mu_J^* e^{\mu_J^*} + J\right)\right]\sigma + \left(\sum_{i=1}^{J} D_i X_{i:m:n} e^{\mu_i} + R_J^* S e^{\mu_J^*}\right)^2$$
$$-\left[\sum_{i=1}^{J} D_i e^{\mu_i} + R_J^* e^{\mu_J^*}\right]\left[\sum_{i=1}^{J} D_i X_{i:m:n}^2 e^{\mu_i} + R_J^* S^2 e^{\mu_J^*}\right] = 0. \tag{37}$$

The above two equations (36) and (37) can be written as

$$(c_1' - c_2' - J)\sigma + d_1' - \mu c_1' = 0 \tag{38}$$

$$A'\sigma^2 + B'\sigma + C' = 0 \tag{39}$$

where

$c_1' = \sum_{i=1}^{J} D_i e^{\mu_i} + R_J^* e^{\mu_J^*}$,

$c_2' = \sum_{i=1}^{J} D_i \mu_i e^{\mu_i} + R_J^* \mu_J^* e^{\mu_J^*}$,

$d_1' = \sum_{i=1}^{J} D_i X_{i:m:n} e^{\mu_i} + R_J^* S e^{\mu_J^*}$,

$d_2' = \sum_{i=1}^{J} D_i X_{i:m:n}^2 e^{\mu_i} + R_J^* S^2 e^{\mu_J^*}$,

$d_3' = \sum_{i=1}^{J} D_i \mu_i X_{i:m:n} e^{\mu_i} + R_J^* \mu_J^* S e^{\mu_J^*}$,

$A' = Jc_1'$, $B' = c_1'\left(d_3' + J\bar{X}\right) - d_1'\left(c_2' + J\right)$,

$C' = d_1^2 - c_1' d_2'$ and $D_i = 1 + R_i$, for $i = 1, \cdots, J$.

The solution to the preceding equations yields the approximate MLE's are

$$\tilde{\mu} = \frac{(c_1^{'} - c_2^{'} - J)\tilde{\sigma} + d_1^{'}}{c_1^{'}} \tag{40}$$

$$\tilde{\sigma} = \frac{-B' + \sqrt{B'^2 - 4A'C'}}{2A'} \tag{41}$$

Consider only positive root of σ; these approximate estimators are equivalent but not unbiased. Unfortunately, it is not possible to compute the exact bias of $\tilde{\mu}$ and $\tilde{\sigma}$ theoretically because of intractability encountered in finding the expectation of $\sqrt{B'^2 - 4A'C'}$.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 547-559

1538 – 9472/09/$95.00

Semi-Parametric of Sample Selection Model Using Fuzzy Concepts

L. Muhamad Safiih A. A. Kamil
University Malaysia Terengganu

M. T. Abu Osman
International Islamic University, Malaysia

The sample selection model has been studied in the context of semi-parametric methods. With the deficiencies of the parametric model, such as inconsistent estimators, semi-parametric estimation methods provide better alternatives. This article focuses on the context of fuzzy concepts as a hybrid to the semi-parametric sample selection model. The better approach when confronted with uncertainty and ambiguity is to use the tools provided by the theory of fuzzy sets, which are appropriate for modeling vague concepts. A fuzzy membership function for solving uncertainty data of a semi-parametric sample selection model is introduced as a solution to the problem.

Key words: Uncertainty, semi-parametric sample selection model, crisp data, fuzzy sets, membership function.

Introduction

The sample selection model has been studied in the context of semi-parametric methods. With the deficiencies of the parametric model, such as inconsistent estimators, etc., semi-parametric estimation methods provide the best alternative to handle the deficiencies. The study of semi-parametric econometrics of the sample selection models has received considerable attention from both statisticians and econometricians in the late of 21st century (Schafgans, 1996). The termed semi-parametric, has been used as a hybrid model for selection models which do not involve parametric forms on error distributions; hence, only the regression function of the model of interest is used. Consideration is based on two perspectives: first, no restriction of estimation of the parameters of interest for the distribution function of the error terms, and second, restricting the functional form of heteroscedasticity to lie in a finite-dimensional parametric family (Schafgans, 1996).

Gallant and Nychka (1987) studied these methods in the context of semi-nonparametric maximum likelihood estimation and applied the method to nonlinear regression with the sample selection model. Newey (1988) used series approximation to the selection correction term which considered regression s-pline and power series approximations. Robinson (1988) focused on the simplest setting of multiple regressions with independent observations, and described extensions to other econometric models, in particular, seemingly unrelated and nonlinear regressions, simultaneous equations, distribution lags and sample selectivity models.

Cosslett (1991) considered semi-parametric estimation of the two-stage method

L. Muhamad Safiih is a statistics/econometrics lecturer in the Faculty of Science and Technology in the Mathematics Department and a Fellow at the Institute of Marine Biotechnology at the University Malaysia Terengganu, 21030 Kuala Terengganu. His research interests are in econometrics modeling, forecasting, applied statistics and fuzzy sets. Email: safiihmd@umt.edu.my. A. A. Kamil is an Associate Professor in Mathematics in the School of Distance Education at the Universiti Sains Malaysia, 11800 USM, Penang, Malaysia. Email: anton@usm.my. M. T. Abu Osman is a Professor in the Kulliyyah of Information and Communication Technology at the International Islamic University, Malaysia, P.O. Box 10, 50728 Kuala Lumpur. His research interests are in combinatorial group theory, fuzzy mathematics, topology and analysis and mathematics educations. Email: abuosman@iium.edu.my.

similar to Heckman (1976) for the bivariate normal case where the first stage consisted of semi-parametric estimation of the binary selection model and the second stage consisted of estimating the regression equation. Ichimura and Lee (1990) proposed an extension of applicability of a semi-parametric approach. It was shown that all models can be represented in the context of multiple index frameworks (Stoker, 1986) and that it can be estimated by the semi-parametric least squares method if identification conditions are met. Andrews (1991) proposed the establishment of asymptotic series estimators for instant polynomial series, trigonometric series and Gallant's Fourier flexible form estimators, for nonparametric regression models and applied a variety of estimands in the regression model under consideration, including derivatives and integrals of the regression function (see also Klein & Spady, 1993; Gerfin, 1996; Vella, 1998; Martin, 2001; Khan & Powell, 2001; Lee & Vella, 2006).

Previous studies in this area concentrated on the sample selection model and used parametric, semi-parametric or nonparametric approaches. None of the studies conducted analyzed semi-parametric sample selection models in the context of fuzzy environment like fuzzy sets, fuzzy logic or fuzzy sets and systems (L. M. Safiih, 2007).

This article introduces a membership function of a sample selection model that can be used to deal with sample selection model problems in which historical data contains some uncertainty. An ideal framework does not currently exist to address problems in which a definite criterion for discovering what elements belong or do not belong to a given set (Miceli, 1998). A fuzzy set, defined by fuzzy sets in a universe of discourse (U) is characterized by a membership function and denoted by the function μ_A, maps all elements of U that take the values in the interval [0, 1] that is $A : X \rightarrow [0, 1]$ (Zadeh, 1965). The concept of fuzzy sets by Zadeh is extended from the crisp sets, that is, the two-valued evaluation of 0 or 1, {0, 1}, to the infinite number of values from 0 to 1, [0, 1]. Brackets { } are used in crisp to indicates sets, whereas square [ ] brackets and parentheses ( ) are used in fuzzy sets to denote real-number closed intervals and open intervals, respectively (see Terano, et al., 1994).

Semi-Parametric Estimation Model

The study of the semi-parametric estimation model involves and considers the two-step estimation approach. The semi-parametric context is a frequently employed method for sample selection models (Vella, 1998) and is a hybrid between the two sides of the semi-parametric approach (i.e., it combines some advantages of both fully parametric and the completely nonparametric). Thus, parts of the model are parametrically specified, while non-parametric estimation issues are used for the remaining part. As a hybrid, the semi-parametric approach shares the advantages and disadvantages of each, in terms that allow a more general specification of the nuisance parameters. In semi-parametric models, the estimators of the parameters of interest are consistent under a broader range of conditions than for parametric models but more precise (converging to the true values at the square root of the sample size) than their nonparametric counterparts.

For a correctly-specified parametric model, estimators for semi-parametric models are generally less efficient than maximum likelihood estimators yet maintain the sensitivity of misspecification for the structural function or other parametric components of the model. In the semi-parametric approach, the differences arise from the weaker assumption of the error term in contrast to the parametric approach. In this study a two-step semi-parametric approach is considered, which generalizes Heckman's two-step procedure. According to Härdle, et al. (1999), Powell (1987) considered a semi-parametric self-selection model that combined the two equation structure of (2.1) with the following weak assumption about the joint distribution of the error terms. For example, the participation equation of the first step is estimated semi-parametrically by the DWADE estimator (Powell, et al., 1989), while applying the Powell (1987) estimator for the second step of the structural equation.

Representation of Uncertainty

Towards representing uncertainty various approaches can be considered. In this

study, the representation of uncertainty identified variables by commonly used approaches, that is, putting a range and a preference function to the desirability of using that particular value within the range. In other words, it is similar to the notion of fuzzy number and membership function which is the function μ_A that takes the values in the interval [0, 1], that is, $A : X \to [0,1]$. For more details about representation of uncertainty, this article concentrates on using fuzzy number and membership function.

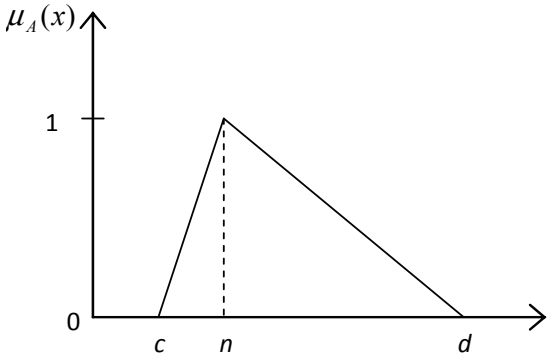
Generally, a fuzzy number represents an approximation of some value which is in the interval terms $[c^{(l)}, d^{(l)}], c^{(l)} \le d^{(l)}$ for l 0, 1, ..., n, and is given by the α- cuts at the α-levels μ_l with $\mu_l = \mu_{l-1} + \Delta\mu, \quad \mu_0 = 0$ and $\mu_n = 1$. A fuzzy number usually provides a better job set to compare the corresponding crisp values. As widely used in practice, each α-cuts ${}^{\alpha}A$ of fuzzy set A are closed and related with an interval of real numbers of fuzzy numbers for all $\alpha \in (0,1]$ and based on the coefficient $A(x)$: if ${}^{\alpha}A \ge \alpha$ then ${}^{\alpha}A = 1$ and if ${}^{\alpha}A < \alpha$ then ${}^{\alpha}A = 0$ which is the crisp set ${}^{\alpha}A$ depends on α.

Closely related with a fuzzy number is the concept of membership function. In this concept, the element of a real continuous number in the interval [0, 1], or a number representing partial belonging or degree of membership are used. Referring to the definition of the membership function, setting the membership grades is open either subjectively to the researcher, depending on his/her intuition, experience and expertise, or objectively based on the analysis of a set of rules and conditions associated with the input data variables. Here, choosing the membership grades is done subjectively, i.e., reflected by a quantitative phenomenon and can only be described in terms of approximate numbers or intervals such as "around 60," "close to 80," "about 10," "approximately 15," or "nearly 50." However, because of the popularity and ease of representing a fuzzy set by the expert - especially when it comes to the theory and applications - the triangular membership function is chosen. It is called a triangular fuzzy number based on a special type of fuzzy number containing three parameters: the grade starts at zero, rises to a maximum and then declines to zero again as the domain increases with its nature; that is, the membership function increases towards the peak and decreasing away from it, and can be represented as a special form as:

$$\mu_A(x) = \begin{cases} \dfrac{(x-c)}{(n-c)} & \text{if } x \in [c,n] \\ 1 & \text{if } x = n \\ \dfrac{(d-x)}{(d-n)} & \text{if } x \in [n,d] \\ 0 & \text{otherwise} \end{cases}$$

The graph of a typical membership function is illustrated in Figure 1.

Figure 1: A Triangular Fuzzy Number



From that function, the α-cuts of a triangular fuzzy number can be define as a set of closed intervals as

$$[(n-c)\alpha + c, (n-d)\alpha + n], \forall \alpha \in (0,1]$$

For the membership function $\mu_A(x)$, the assumptions are as follows:

(i) monotonically increasing function for membership function $\mu_A(x)$ with $\mu_A(x) = 0$ and $\lim_{x\to\infty} \mu_A(x) = 1$ for $x \le n$.

(ii) monotonically decreasing function for membership function $\mu_A(x)$ with $\mu_A(x)=1$ and $\lim_{x\to\infty}\mu_A(x)=0$ for $x \geq n$.

The α-cuts and LR Representation of a Fuzzy Number

Prior to delving into fuzzy modeling of PSSM, an overview and some definitions used in this article are presented (Yen, et al., 1999; Chen & Wang, 1999); the definitions and their properties are related to the existence of fuzzy set theory and were introduced by Zadeh (1965).

Definition*:* the fuzzy function is defined by $f: X \times \widetilde{A} \to \widetilde{Y}; \widetilde{Y} = f(x, \widetilde{A})$, where

1. $x \in X$; X is a crisp set, and
2. $\widetilde{A}$ is a fuzzy set, and
3. $\widetilde{Y}$ is the co-domain of x associated with the fuzzy set $\widetilde{A}$.

Definition*:* Let $A \in F(\Re)$ be called a fuzzy number if:

1) $x \in \Re$ such that $\mu_A(x)=1$,
2) for any $\alpha \in [0,1]$, and
3) $A_\alpha = [x, \mu_{A_\alpha}(x) \geq a]$, is a closed interval with $F(\Re)$ representing all fuzzy sets, $\Re$ is the set of real numbers.

Definition*:* a fuzzy number A on $\Re$ is defined to be a triangular fuzzy number if its membership function $\mu_A(x): \Re \to [0,1]$ is equal to

$$\mu_A(x) = \begin{cases} \dfrac{(x-l)}{(m-l)} & if\ x \in [l,m] \\ 1 & if\ x = m \\ \dfrac{(u-x)}{(u-m)} & if\ x \in [m,u] \\ 0 & otherwise \end{cases}$$

where $l \leq m \leq u$, x is a model value with l and u be a lower and upper bound of the support of A respectively. Thus, the triangular fuzzy number is denoted by (l, m, u). The support of A is the set elements $\{x \in \Re \mid l < m < u\}$. A non-fuzzy number by convention occurs when $l = m = u$.

Theorem 1:

The values of estimator coefficients of the participation and structural equations for fuzzy data converge to the values of estimator coefficients of the participation and structural equations for non-fuzzy data respectively whenever the value of $\alpha - \text{cut}$ tends to 1 from below.

Proof:

From the centroid method followed to obtain the crisp value, the fuzzy number for all observation of w_i is

$$W_{ic} = \frac{1}{3}\left(Lb(w_i) + w_i + Ub(w_i)\right)$$

when there is no $\alpha - cut$. The lower bound and upper bound for each observation is referred to by the definition above.

Because the triangular membership function is followed (see Figure 2) then

$$A = \left(Lb(w_{i(\alpha)}), \alpha\right) \text{ and } B = \left(Ub(w_{i(\alpha)}), \alpha\right),$$

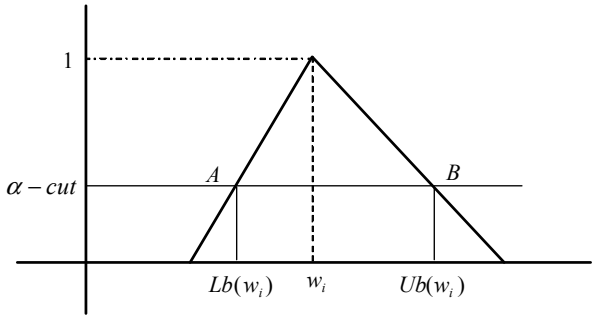
where

$$Lb(w_{i(\alpha)}) = Lb(w_i) + \alpha\left(w_i - Lb(w_i)\right)$$

and

$$Ub(w_{i(\alpha)}) = Ub(w_i) + \alpha\left(w_i - Ub(w_i)\right)$$

Figure 2: Membership Function and $\alpha - \text{cut}$



Applying the $\alpha - cut$ into the triangular membership function, the fuzzy number obtained depending on the given value of the $\alpha - cut$ over the range 0 and 1 is as follows:

$$W_{ic(\alpha)} = \frac{\begin{array}{c} Lb(w_i) + \alpha\left(w_i - Lb(w_i)\right) \\ +w_i + Ub(w_i) + \alpha\left(w_i - Ub(w_i)\right) \end{array}}{3} = \frac{Lb(w_{i(\alpha)}) + w_i + Ub(w_{i(\alpha)})}{3}.$$

When α approaches 1 from below then $Lb(w_{i(\alpha)}) \to w_i$ and $Ub(w_{i(\alpha)}) \to w_i$, and is obtained as

$$W_{ic(\alpha)} \to \frac{w_i + w_i + w_i}{3} = w_i \ ,$$

$$W_{ic(\alpha)} \to w_i .$$

Thus, when α approaches 1 from below, then $W_{ic(\alpha)} \to w_i$. Similarly, for all observations x_i and z_i, $X_{ic(\alpha)} \to x_i$ and $Z_{ic(\alpha)} \to z_i$ respectively, as α tends to 1 from below. This implies that the values of estimator coefficients of the participation and structural equations for fuzzy data converge to the values of estimator coefficients of the participation and structural equations for non-fuzzy data respectively whenever the value of $\alpha - \text{cut}$ tend to 1 from below

Definition*:* An LR-type fuzzy number denoted as $\widetilde{Y}$ with functions $L(Y) = f_1((\frac{1}{\beta})(Y_C - Y))$ and $R(Y) = f_2((\frac{1}{\gamma})(Y - Y_C))$. $\widetilde{Y}$ consists of the lower bound (Y_L), center (Y_C) and upper bound (Y_U). Satisfying

$$L(Y_L) = R(Y_U) = 0(\alpha_{\min})$$

and

$$L(Y_C) = R(Y_C) = 1(\alpha_{\max}) .$$

The size of $\widetilde{Y}$ is $Y_U - Y_L$ where $\alpha_{\min}$ and $\alpha_{\max}$ can be any predetermined levels.

Theorem 2:

If an LR-type fuzzy number is denoted as $\widetilde{Y}'$ with $L(Y')$ and $R(Y')$ functions of $f_1((\frac{1}{k_1\beta})(Y_C^{'} - Y'))$ and $f_2((\frac{1}{k_2\beta})(Y' - Y_C^{'}))$ respectively, then, (Y_L), (Y_C) and (Y_U) of $\widetilde{Y}'$ are

$$Y_C^{'} - k_1(Y_C - Y_L), Y_C^{'}$$

and

$$Y_C^{'} + k_2(Y_U - Y_C) .$$

Proof:

Because for $\widetilde{Y}$

$$L(Y_L) = f_1\left(\frac{1}{\beta}(Y_C - Y_L)\right) = R(Y_U) = f_2\left(\frac{1}{\gamma}(Y_U - Y_C) = 0\right) ,$$

$$L(Y_C) = f_1(0) = R(Y_C) = f_2(0) = 1 ,$$

then, for $\widetilde{Y}'$

$$L(Y_C^{'} - k_1(Y_C - Y_L)) = f_1\left(\frac{1}{k_1\beta}(Y_C^{'} - Y_C^{'} + k_1(Y_C - Y_L))\right) = f_1\left(\frac{1}{\beta}(Y_C - Y_L)\right) = 0$$

and

$$R(Y_C^{'} - k_2(Y_U - Y_C)) = f_2\left(\frac{1}{k_2\gamma}(Y_C^{'} + k_2(Y_U - Y_C) - Y_C^{'})\right) = f_2\left(\frac{1}{\gamma}(Y_U - Y_C)\right) = 0$$

$$L(Y_C^{'}) = f_1(0) = R(Y_C^{'}) = f_2(0) = 1$$

Thus, Theorem 2 is proven.

Methodology

Development of Fuzzy Semi-Parametric Sample Selection Models

Prior to constructing a fuzzy SPSSM, the sample selection model purpose by Heckman (1976) is considered. In SPSSM, it is assumed that the distributional assumption of $(\varepsilon_i,\ u_i)$ is weaker than the distributional assumption of the parametric sample selection model. The distributional assumption that exists in Heckman (1979) model is more stringent than anything else. However, the Heckman (1979) estimator becomes inconsistent if the assumption is violated. Härdle, et al. (1999) highlighted that ample reason exists to develop consistent estimators for PSSM with weaker distributional assumptions. Thus, the sample selection model is now called a semi-parametric of sample selection model (SPSSM).

In the development of SPSSM modeling using the fuzzy concept, as a development of fuzzy PSSM, the basic configuration of fuzzy modeling is still considered as previously mentioned (i.e., involved fuzzification, fuzzy environment and defuzzification). For the fuzzification stage, an element of real-valued input variables is converted in the universe of discourse into values of a membership fuzzy set. At this approach, a triangular fuzzy number is used over all observations. The α-cut method with an increment value of 0.2 started with 0 and increases to 0.8. This is then applied to the triangular membership function to obtain a lower and upper bound for each observation ($x_i,\ w_i$ and z_i^*), defined as:

$$\tilde{w}_{isp} = (w_{il}, w_{im}, w_{iu}),\ \tilde{x}_{isp} = (x_{il}, x_{im}, x_{iu})$$

and

$$\tilde{z}^*_{isp} = (z_{il}, z_{im}, z_{iu}).$$

Following their memberships functions, respectively defined, results in the following forms:

$$\mu_{\tilde{w}_i^* sp}(z) = \begin{cases} \dfrac{(w - w_{il})}{(w_{im} - w_{il})} & if\ w \in [w_{im}, w_{im}] \\ 1 & if\ w = w_{im} \\ \dfrac{(w_{iu} - w_{im})}{(w_{iu} - w_{im})} & if\ w \in [w_{im}, w_{iu}] \\ 0 & otherwise \end{cases}$$

$$\mu_{\tilde{x}_i sp}(x) = \begin{cases} \dfrac{(x - x_{il})}{(x_{im} - x_{il})} & if\ x \in [x_{il}, x_{im}] \\ 1 & if\ x = x_{im} \\ \dfrac{(x_{iu} - x)}{(x_{iu} - x_{im})} & if\ x \in [x_{im}, x_{iu}] \\ 0 & otherwise \end{cases}$$

and

$$\mu_{\tilde{z}_i sp}(z) = \begin{cases} \dfrac{(z - z_{il})}{(z_{im} - z_{il})} & if\ z \in [z_{im}, z_{im}] \\ 1 & if\ z = z_{im} \\ \dfrac{(z_{iu} - z_{im})}{(z_{iu} - z_{im})} & if\ z \in [z_{im}, z_{iu}] \\ 0 & otherwise \end{cases}$$

In order to solve the model in which uncertainties occur, fuzzy environments such as fuzzy sets and fuzzy numbers are more suitable as the processing of the fuzzified input parameters. Because, it is assumed that some original data contains uncertainty, under the vagueness of the original data, the data can be considered as fuzzy data. Thus, each observation considered has variation values. The upper and lower bounds of the observation are commonly chosen based on the data structure and experience of the researchers. For a large-sized observation, the upper and lower bounds of each observation are difficult to obtain.

Based on the fuzzy number, a fuzzy SPSSM is built with the form as:

$$\tilde{z}^{*}_{i_{sp}} = \tilde{w}^{'}_{i_{sp}}\gamma + \tilde{\varepsilon}_{i_{sp}} \qquad i = 1,...,N$$

$$d_i = 1 \;\; if \;\; d^{*}_i = \tilde{x}^{'}_{i_{sp}}\beta + \tilde{u}_{i_{sp}} > 0,$$

$$d_i = 0 \; otherwise \; i = 1,...,N$$

$$z_i = z^{*}_{ic_{sp}} d_i$$

The terms $\tilde{w}_{i_{sp}}$, $\tilde{x}_{i_{sp}}$, $\tilde{z}^{*}_{i_{sp}}$, $\tilde{\varepsilon}_{i_{sp}}$ and $\tilde{u}_{i_{sp}}$ are fuzzy numbers with the membership functions $\mu_{\tilde{W}_{isp}}$, $\mu_{\tilde{X}_{i_{sp}}}$, $\mu_{\tilde{Z}_{isp}}$, $\mu_{\tilde{\varepsilon}_{isp}}$ and $\mu_{\tilde{u}_{isp}}$, respectively. Because the distributional assumption for the SPSSM is weak, for the analysis of the fuzzy SPSSM it is also assumed that the distributional assumption is weak.

To determine an estimate for γ and β of the fuzzy parametric of a sample selection model, one option is to defuzzify the fuzzy observations $\tilde{W}^{'}_{i_{sp}}$, $\tilde{X}^{'}_{i_{sp}}$ and $\tilde{Z}^{*}_{i_{sp}}$. This means converting the triangular fuzzy membership real-value into a single (crisp) value (or a vector of values) that, in the same sense, is the best representative of the fuzzy sets that will actually be applied. The centroid method or the center of gravity method is used to compute the outputs of the crisp value as the center of the area under the curve. Let $W_{ic_{sp}}$, X_{icsp}, and $Z^{*}_{ic_{sp}}$ be the defuzzified values of $\tilde{W}_{i_{sp}}$, $\tilde{X}_{i_{sp}}$, and $\tilde{Z}^{*}_{i_{sp}}$ respectively. The calculation of the centroid method for $W_{ic_{sp}}$, $X_{ic_{sp}}$, and $Z^{*}_{ic_{sp}}$ respectively is via the following formulas:

$$W_{ic_{sp}} = \frac{\int_{-\infty}^{\infty} w\mu_{\tilde{w}_i}(w)dw}{\int_{-\infty}^{\infty} \mu_{\tilde{w}_i}(w)dw} = \frac{1}{3}(W_{i_l} + W_{i_m} + W_{i_u}),$$

$$X_{ic_{sp}} = \frac{\int_{-\infty}^{\infty} x\mu_{\tilde{x}_i}(x)dx}{\int_{-\infty}^{\infty} \mu_{\tilde{x}_i}(x)dx} = \frac{1}{3}(X_{i_l} + X_{i_m} + X_{i_u}),$$

and

$$Z^{*}_{ic_{sp}} = \frac{\int_{-\infty}^{\infty} z\mu_{\tilde{z}_i}(z)dz}{\int_{-\infty}^{\infty} \mu_{\tilde{z}_i}(z)dz} = \frac{1}{3}(Z_{i_l} + Z_{i_m} + Z_{i_u}).$$

Thus, the crisp values for the fuzzy observation are calculated following the centroid formulas as stated above. To estimate γ_{sp} and β_{sp} with the SPSSM approach, applying the procedure as in Powell, then the parameter is estimated for the fuzzy semi-parametric sample selection model (fuzzy SPSSM). Before obtaining a real value for the fuzzy SPSSM coefficient estimate, first the coefficient estimated values of γ and β are used as a shadow of reflection to the real one. The values of $\hat{\gamma}$ and $\hat{\beta}$ are then applied to the parameters of the parametric model to obtain a real value for the fuzzy SPSSM coefficient estimates of γ_{sp}, β_{sp}, $\sigma_{\varepsilon_{isp}}$, $u_{i_{sp}}$. The Powell (1987) SPSSM procedure is then executed using the XploRe software.

The Powell SPSSM procedure combines the two-equation structure as shown above but has a weaker assumption about the joint distribution of the error terms:

$$f(\varepsilon_{i_{sp}}, u_{i_{sp}} \mid w_{i_{sp}}) = f(\varepsilon_{i_{sp}}, u_{i_{sp}} \mid w^{'}_{i_{sp}}\gamma).$$

For this reason, it is assumed that the joint densities of $\varepsilon_{i_{sp}}$, $u_{i_{sp}}$ (conditional on $w_{i_{sp}}$) are smooth but unknown functions $f(\cdot)$ that depend on $w_{i_{sp}}$ only through the linear model $w^{'}_{i_{sp}}\gamma$. Based on this assumption, the regression function for the observed outcome z_i takes the following form:

$$\begin{aligned} E(z_i \mid x_{i_{sp}}) &= E(z^{*}_{i_{sp}} \mid w_{i_{sp}}, d^{*}_{i_{sp}} > 0) \\ &= w^{'}_{i_{sp}}\gamma + E(u_{i_{sp}} \mid w_{i_{sp}}, x^{'}_{i_{sp}}\beta > -\varepsilon_{i_{sp}}) \\ &= w^{'}_{i_{sp}}\gamma + \lambda(x^{'}_{i_{sp}}\beta) \end{aligned}$$

where $\lambda(\cdot)$ is an unknown smooth function. The Powell idea of SPSSM is based upon two observations, i and j, with conditions $w_{i_{sp}} \neq w_{j_{sp}}$ but $w_{i_{sp}}^{'}\gamma = w_{j_{sp}}^{'}\gamma$. With this condition, the unknown function $\lambda(\cdot)$ can be differenced out by subtracting the regression functions for i and j:

$$\begin{aligned} E(z_{i_{sp}}^{*} \mid w = w_{i_{sp}}) - E(z_{j_{sp}}^{*} \mid w = w_{j_{sp}}) \\ = (w_{i_{sp}} - w_{j_{sp}})'\gamma + \lambda(x_{i_{sp}}^{'}\beta) - \lambda(x_{j_{sp}}^{'}\beta) \\ = (w_{i_{sp}} - w_{j_{sp}})'\gamma \end{aligned}$$

This is the basic idea underlying the γ estimator proposed by Powell (1987). Powell's procedure is from the differences, regress z_i on differences in $w_{i_{sp}}$, as the concept of closeness with two estimated indices $x_{i_{sp}}^{'}\hat{\beta}$ and $x_{j_{sp}}^{'}\hat{\beta}$ (hence $\lambda(x_{i_{sp}}^{'}\hat{\beta}) - \lambda(x_{j_{sp}}^{'}\hat{\beta}) \approx 0$). Thus, γ can be estimated by a weighted least squares estimator:

$$\hat{\gamma}_{Powell} = \left[\binom{n}{2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\hat{\varpi}_{ij}N(w_{i_{qp}} - w_{j_{sp}})(w_{i_{qp}} - w_{j_{qp}})'\right]^{-1} \times \left[\binom{n}{2}^{-1}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\hat{\varpi}_{ij}N(w_{i_{qp}} - w_{j_{qp}})(z_{i_{qp}} - z_{j_{qp}})\right]$$

Where weights $\hat{\varpi}_{ij}N$ are calculated by $\hat{\varpi}_{ij}N = \frac{1}{h}\kappa\left(\frac{x_{i_{sp}}^{'}\hat{\beta} - x_{j_{sp}}^{'}\hat{\beta}}{h}\right)$ with a symmetric kernel function $\kappa(\cdot)$ and bandwidth h. As shown in earlier equations, this tacitly assumes that $\hat{\beta}$ has previously been obtained as an estimate β. Based on this assumption, a single index model is obtained for the decision equation in place of the probit model (probit step) in the parametric case:

$$P(d_i(d_i^{'} > 0 \mid x) = 1) = g(x_i^{'}\beta)$$

where $g(\cdot)$ is an unknown, smooth function. Using this and given $\hat{\beta}$, the second step consists of estimating γ. Executing the Powell procedure by XploRe software takes the data as input from the outcome equation (x and y, where x may not contain a vector of ones). The vector *id* containing the estimate for the first-step index $x_{i_{sp}}^{'}\hat{\beta}$, and the bandwidth vector *h* where *h* is the threshold parameter *k* that is used for estimating the intercept coefficient from the first element. The bandwidth *h* from the second element (not covered in this study) is used for estimating the slope coefficients. For fuzzy PSSM, the above procedure is followed, and then another set of crisp values $W_{ic_{sp}}$, $X_{ic_{sp}}$ and $Z_{ic_{sp}}$ are obtained. Applying the α-cut values on the triangular membership function of the fuzzy observations $\widetilde{W}_{i_{sp}}$, $\widetilde{X}_{i_{sp}}$ and $\widetilde{Z}_{i_{sp}}$ with the original observation, fuzzy data without α-cut and fuzzy data with α-cut to estimate the parameters of the fuzzy SPSSM. The same procedure above is applied. The parameters of the fuzzy SPSSM are estimated. From the various fuzzy data, comparisons will be made on the effect of the fuzzy data and α-cut with original data on the estimation of the SPSSM.

Data Description

The data set used for this study is from the 1994 Malaysian Population and Family Survey (MPFS-94). This survey was conducted by National Population and Family Development Board of Malaysia under the Ministry of Women, Family and Community Development Malaysia. The survey was specifically for married women, providing relevant and significant information for the problem of married womens' status regarding wages, educational attainment, household composition and other socioeconomic characteristics. The original MPFS-94 sample data comprised 4,444 married women. Based on the sequential criteria (Mroz, 1984) the analyses were limited to the completed information provided by married women; in addition, respondents whose information was incomplete

(for example, no recorded family income in 1994, etc.), were removed from the sample.

The resulting sample data consisted of 1,100 married women, this accounted for 39.4% who were employed, the remaining 1,692 (60.6%) were considered as non-participants. The data set used in this study consisted of 2,792 married women. Selection rules (Martins, 2001) were applied to create the sample criteria for selecting participant and non participant married women on the basis of the MPFS-94 data set, which are as follows:

a) Married and aged below 60;
b) Not in school or retired;
c) Husband present in 1994; and
d) Husband reported positive earnings for 1994.

Study Variables

Following the literature (see Gerfin, 1996; Martins, 2001; Christofides, et al., 2003), the model employed in this study consists of two equations or parts. The first equation - the probability that a married women participates in the labor market - is the so-called participation equation. Independent variables involved are: AGE (age in years divided by 10), AGE2 (age squared divided by 100), EDU (years of education), CHILD (the number of children under age 18 living in the family), HW (log of husband's monthly wage). The standard human capital approach was followed for the determination of wages, with the exception of potential experience. The potential experience variable in the data set was calculated by age-edu-6 rather than actual work experience. In order to manage these problems a method advanced by Buchinsky (1998) was used. Instead of considering the term $Q_w = \xi_1 EXP + \xi_2 EXP^2$ in the wage equation i.e., *EXP* is the unobserved actual experience, we use the alternative for women's time is child rearing and the home activities related to child rearing, then the specification of Q_z given by:

$$Q_z = \gamma_1 PEXP + \gamma_2 PEXP2 + \gamma_3 PEXPCHD + \gamma_4 PEXPCHD2$$

The second equation called the wage equation. The dependent variable used for the analysis was the log hourly wages (z). While, the independent variables were EDU, PEXP (potential work experience divided by 10), PEXP2 (potential experience squared divided by 100), PEXPCD (PEXP interacted with the total number of children) and PEXPCHD2 (PEXP2 interacted with the total number of children). Both the participation and wage equations were considered as the specification I and II respectively, that is, the most basic one of SSM.

According to Kao and Chin (2002), the regression parameters (β, γ) should be estimated from the sample data and, if some of the observations in the equation X_{ij} and Y_i are fuzzy, then they fall into the category of fuzzy regression analysis. For the data used in this study, it was assumed that uncertainty was present, therefore, instead of crisp data, fuzzy data are more appropriate. In the participation equation, fuzzy data was used for the independent variables (x): AGE (age in year divided by 10), AGE2 (age square divided by 100) HW (log of husband's monthly wage). For the wage equation, fuzzy data used for the dependent variable was the log hourly wages (z) while the independent variables (x) for fuzzy data involved the variables PEXP (potential work experience divided by 10), PEXP2 (potential experience squared divided by 100), PEXPCD (PEXP interacted with the total number of children) and PEXPCHD2 (PEXP2 interacted with the total number of children). In our study, the observations in the fuzzy participation and fuzzy wage equations involved fuzzy and non-fuzzy data, i.e. a mixture between fuzzy and non-fuzzy data, thus the variables fall into the category of fuzzy data (Kao and Chyu, 2002). For instance, the exogenous variables *AGE, AGE*2 and *HWS* in the participation and the variables *PEXP, PEXP*2*, PEXPCHD* and *PEXPCHD*2 in the wage equations are in the form of fuzzy data. These fuzzy exogenous variables are denoted as $A\tilde{G}E$, $A\tilde{G}E2$ $H\tilde{W}S$ and $P\tilde{E}XP$, $PE\tilde{X}P2$, $PEX\tilde{P}CHD$, $PEX\tilde{P}CHD2$, respectively. In accord with general sample selection model, the exogenous variables *EDU* and *CHILD* in the participation

and the exogenous variable *EDU* in the fuzzy wage equation are considered as non-fuzzy data. However *EDU* and *CHILD* are considered as fuzzy data.

Results

A semi-parametric estimation obtained due to the so-called curse of dimensionality and asymptotic distribution is unknown. Here the results that applied to the most basic estimators are presented; that is, the participant and wage equation of the DWADE estimator and the Powell estimator, respectively. Both estimators are consistent with $\sqrt{n}-$consistency and asymptotic normality.

Participation Equation in the Wage Sector

The participation equation using the DWEDE estimator is presented in Table 1 along with FSPSSM results for comparison purposes. The first column used the DWADE estimator with bandwidth values h = 0.2 without the constant terms. The DWADE estimator shares the ADE estimator of the semi-parametric sample selection model (SPSSM). This is followed by the fuzzy semi-parametric sample selection model (FPSSM) with $\alpha-$cuts 0.0, 0.2, 0.4, 0.6 and 0.8 respectively. At first the estimate coefficient suggests that all variables except AGE are significant (significantly and negatively estimated coefficient on AGE2 and CHILD, while a positive and significant coefficient was estimated for EDU and HW). However, only CHILD shows a statistically significant effect at the 5% level – an unexpected and important result. Although in the conventional parametric model, it appears together with EDU, in the context of SPSSM, only estimates of the CHILD effect appears to be significantly relevant, which is more aligned with economic theory.

For comparison purposes, the FSPSSM was used. The estimated coefficient gives a similar trend with the SPSSM (i.e., significant for variables AGE2, EDU, CHILD and HW). The results show a significant and positive coefficient estimate for EDU and HW, and a significant but negative estimated coefficient on AGE2 and CHILD. In the FSPSSM context, the CHILD coefficient appears to be statistically significant at the 5% level. This is an interesting finding and it should be pointed out that using this approach the standard errors for the parameter were much smaller when compared to those in conventional SPSSM. This provides evidence that this approach is better in estimating coefficients and provides a considerable efficiency gain compared to those in the conventional semi-parametric model. In addition, the coefficient estimated from FSPSSM was considerably close to the coefficient estimated with conventional SPSSM. Hence, the coefficient estimated from FSPSSM is consistent even though it involves uncertain data.

The Wage Equation in the Wage Sector

The wage equation using the Powell estimator of SPSSM is presented in Table 2 with FSPSSM results for comparison purposes. The first column used the Powell estimator with bandwidth values h = 0.2 without the constant terms. The other columns show results given by the fuzzy semi-parametric sample selection model (FPSSM) with $\alpha-$cuts 0.0, 0.2, 0.4, 0.6 and 0.8 respectively.

At first the coefficient estimate suggested that the whole variable was significant (significant and negatively estimated coefficient on EDU, PEXP2 and PEXPCHD, while a positive and significant coefficient was estimated for PEXP and PEXPCHD2). As the estimated coefficient, the results for whole variable statistical significance at the 5% level resulted in a significant result. The results reveal significant differences between the SPSSM compared to the PSSM method of correcting sample selectivity bias. This increased the results obtained in SPSSM where not all variables in PSSM contributed significantly regarding married women involved in wage sectors.

For comparison purposes it was then applied with the FSPSSM. The estimated coefficient was significant for all variables. The results show significant and positive coefficient estimates for PEXP and PEXPCHD2, significant but negative estimated coefficients on EDU, PEXP2 and PEXPCHD. The coefficient for all variables appears to be relevant with statistical

significance at the 5% level. It should be noted that, the standard errors for the parameter EDU, PEXP and PEXP2 were much smaller when compared to those in the conventional SPSSM. This provides evidence that this method is considerably more efficient than the conventional semi-parametric model. The coefficient estimated obtained from FSPSSM is also considerably close to the coefficient estimated via conventional SPSSM. In other words, applying FSPSSM, the coefficient estimated is consistent even though the data may contain uncertainties.

Conclusion

For comparison purposes of the participant equation, the estimated coefficient and significant factor gives a similar trend as the SPSSM. However, an interesting finding and the most significant result appears by applying the

Table1: Semi-Parametric and Fuzzy Semi-Parametric Estimates for the Participation Equation

Participation Equation	Coefficients					
	DWADE	Fuzzy Selection Model				
		α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 0.0
AGE	−0.002048 (1.233)	−0.0015393 (1.150)	−0.0043978 (1.151)	−0.0015934 (1.234)	−0.0016184 (1.232)	−0.001642 (1.232)
AGE2	−0.00016099 (0.1754)	−0.00016584 (0.1624)	−0.00020722 (0.1627)	−0.00016629 (0.1763)	−0.00016651 (0.1765)	−0.00016673 (0.1767)
EDU	0.00034766 (0.02116)	0.00023044 (0.02015)	0.00011323 (0.02015)	0.00023044 (0.02115)	0.00023044 (0.02062)	0.00023044 (0.02062)
CHILD	−0.0039216* (0.06573)	−0.0044301* (0.06341)	−0.0048986* (0.0634)	−0.0044301* (0.06571)	−0.0044301* (0.06485)	−0.0044301* (0.06484)
HW	0.044008 (0.1632)	0.050262 (0.1402)	0.05597 (0.1396)	0.049549 (0.1485)	0.049189 (0.1437)	0.048832 (0.1432)

*5% level of significance

Table 2: Semi-Parametric and Fuzzy Semi-Parametric Estimates for the Wage Equation

Wage Equation	Coefficients					
	Powell	Fuzzy Selection Model				
		α = 0.8	α = 0.6	α = 0.4	α = 0.2	α = 0.0
EDU	−0.0112792 (0.005262)	−0.0109003 (0.005258)	−0.010939 (0.005258)	−0.011346 (0.005259)	−0.011385 (0.005259)	−0.0114256 (0.005258)
PEXP	0.544083* (0.1099)	0.540864* (0.1096)	0.538776* (0.1094)	0.534385* (0.1093)	0.532247* (0.1092)	0.530069* (0.109)
PEXP2	−0.160272* (0.02633)	−0.159762* (0.0263)	−0.159524* (0.0263)	−0.158781* (0.02632)	−0.158525* (0.02632)	−0.158259* (0.02632)
PEXPCHD	−0.161205* (0.02453)	−0.159863* (0.02453)	−0.159583* (0.02455)	−0.15889* (0.02459)	−0.158584* (0.02461)	−0.158262* (0.02463)
PEXPCHD2	0.046591* (0.008485)	0.0463242* (0.008485)	.0462221* (0.008493)	0.0458118* (0.008508)	.0457004* (0.008511)	0.0455835* (0.008517)

*5% level of significance

FPSSM, that is, the FSPSSM is a better estimate when compared to the SPSSM in terms of the standard error of the coefficient estimate. The standard errors of the coefficient estimate for the FSPSSM are smaller when compared to the conventional SPSSM. This is evidence that the FSPSSM approach is much better in estimate coefficient and results in considerable efficiency gain than the conventional semi-parametric model. The coefficient estimate obtained was also considerably close to the coefficient estimate of conventional SPSSM, hence providing evidence that the coefficient estimate is consistent even when data involves uncertainties.

The wages equation is similar to the PSSM in terms of the coefficient estimation and significance factors. However, applying the FPSSM resulted in the most significant results when compared to the PSSM, the coefficient estimates of most variables had small standard errors. The rest is considerably close to the standard error of SPSSM. As a whole, the FSPSSM gave a better estimate compared to the SPSSM. In terms of consistency the coefficient estimate for all variables of FSPSSM were not much different to the coefficient estimates of SPSSM even though the values of the $\alpha-\text{cuts}$ increased (from 0.0 to 0.8). In the other words, by observing the coefficient estimate and consistency, fuzzy model (FPSSM) performs much better than the model without fuzzy (PSSM) for the wage equation.

References

Andrews, D. W. K. (1991). Asymptotic normality of series estimation for nonparametric and semiparametric regression models. *Econometrica*, *59*, 307-345.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics*, *13*, 1-30.

Chen, T., & Wang, M. J. J. (1999). Forecasting methods using fuzzy concepts. *Fuzzy Sets and Systems*, *105*, 339-352.

Christofides, L. N., Li, Q., Liu, Z., & Min, I. (2003). Recent two-stage sample selection procedure with an application to the gander wage gap. *Journal of Business and Economic Statistics*, *21*(3), 396-405.

Cosslett, S. (1991). Semiparametric estimation of a regression models with sample selectivity. In Barnett, W. A., Powell, J., & Tauchen, G. E. (*Eds.*), 175-198. *Nonparametric and semiparametric estimation methods in econometrics and statistics*. Cambridge, MA: Cambridge University Press.

Gallant, R., & Nychka, D. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*, 363-390.

Gerfin, M. (1996). Parametric and semiparametric estimation of the binary response model of labor market participation. *Journal of Applied Econometrics*, *11*, 321-339.

Härdle, W., Klinke, S., & Müller, M. (1999). Xplore learning guide. Berlin: Springer-Verlag.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimation for such models. *Annals of Economic and Social Measurement*, *5*, 475-492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153-161.

Ichimura, H., & Lee, L. F. (1990). Semiparametirc least square estimation of multiple index models: Single equation estimation. In Barnett, W. A., Powell, J., & Tauchen, G. E. (*Eds.*), 175-198. *Nonparametric and semiparametric estimation methods in econometrics and statistics*. Cambridge, MA: Cambridge University Press.

Kao, C., & Chin, C. L. (2002). A fuzzy linear regression model with better explanatory power. *Fuzzy Sets and Systems*, *126*, 401-409.

Khan, S., & Powell, J. L. (2001). Two-step estimation of semiparametric censored regression models. *Journal of Econometrics*, *103*, 73-110.

Klein, R., & Spady, R. (1993). An efficient semiparametric estimator of the binary response model. *Econometrica*, *61*(*2*), 387-423.

Lee, M-J., & Vella, F. (2006). A semi-parametric estimator for censored selection models with endogeneity. *Journal of Econometrics*, *130*, 235-252.

Martin, M. F. O. (2001). Parametric and semiparametric estimation of sample selection models: An empirical application to the female labor force in Portugal. *Journal of Applied Econometrics*, *16*, 23-39.

Miceli, D. (1998). *Measuring poverty using fuzzy sets*. Discussion paper no.38; National centre for social and economic modeling, University of Canberra.

Mroz, T. A. (1984). *The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions*. Ph.D. dissertation, Stanford University.

Newey, W. (1988). *Two step series estimation of sample selection models*. Department of Economics, MIT working paper no. E52 - 262D, 1-17.

Powell, J. L. (1987). Semi-parametric estimation of bivariate latent variable models. Social Systems Research Institute. University of Wisconsin-Madison, Working paper No.8704.

Powell, J., Stock, J. H., & Stoker, T. M. (1989). Semi-parametric estimation of index coefficients. *Econometrica*, *57*, 1403-1430.

Safiih, L. M. (2007). *Fuzzy semi-parametric of a sample selection model*. Ph.D. dissertation. University Science of Malaysia.

Robinson, P. M. (1988). Root-N consistent semi-parametric regression. *Econometrica*, *56*, 931-954.

Schafgans, M. (1996). *Semiparametric estimation of a sample selection model: Estimation of the intercept; theory and applications*. Ph. D. dissertation, Yale University.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, *54*, 1461-1481.

Terano, T., Asai, K., & Sugeno, M. (1994). *Applied fuzzy systems*. Cambridge, MA: AP Professional.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, *33*, 127-169.

Yen, K. K., Ghoshray. S., & Roig. G. (1999). A linear regression model using triangular fuzzy number coefficients. *Fuzzy Sets and Systems*, *106*, 167-177.

Zadeh, L. A. (1965). Fuzzy Sets and Systems. In Fox, J. (*Ed.*), *System Theory*. Brooklyn, NY: Polytechnic Press.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 560-565

1538 – 9472/09/$95.00

Performance Ratings of an Autocovariance Base Estimator (ABE) in the Estimation of GARCH Model Parameters When the Normality Assumption is Invalid

Daniel Eni
Federal University of Petroleum
Resources, Effurun- Nigeria

The performance of an autocovariance base estimator (ABE) for GARCH models against that of the maximum likelihood estimator (MLE) if a distribution assumption is wrongly specified as normal was studied. This was accomplished by simulating time series data that fits a GARCH model using the Log normal and t-distributions with degrees of freedom of 5, 10 and 15. The simulated time series was considered as the true probability distribution, but normality was assumed in the process of parameter estimations. To track consistency, sample sizes of 200, 500, 1,000 and 1,200 were employed. The two methods were then used to analyze the series under the normality assumption. The results show that the ABE method appears to be competitive in the situations considered.

Key words: Autocovariance Functions, Parameter Estimation, Garch, Normality.

Introduction

The assumption of constant variance in the traditional time series models of ARMA is a major impediment to their applications in financial time series data where heteroscedasticity is obvious and cannot be ignored. To solve this problem, Engle (1982) proposed the Autoregressive Conditional Heteroscedascity (ARCH) model. In his first application, however, Engle noted that a high order of ARCH is needed to satisfactorily model time varying variances and that many parameters in ARCH will create convergence problems for maximization routines. To address these difficulties, Bollerslev (1986) extended Engle's model, developing the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model. GARCH models time-varying variances as a linear function of past square residuals and of its past value. It has proved useful in interpreting volatility clustering effects and has gained wide acceptance in measuring the volatility of financial markets. The ARCH and GARCH models are both known as symmetric models.

Other extensions based on observed characteristics of financial time series data exist and include some asymmetric models. Examples of asymmetric models are Nelson's (1991) exponential GARCH (EGARCH) model, Glosten, Jaganathan and Runkle's (1993) GJR-GARCH and Zakoian's (1994) threshold model (T GARCH). These model and interpret leverage effects, where volatility is negatively correlated with returns. In addition, the Fractionally Integrated GARCH model (FIGARCH) (Baillie, Bollerslev & Mikeson, 1996) was introduced to model long memory via the fractional operator $(1\text{-}L)^d$, and the GARCH in mean model allows the mean to influence the variance.

These models are popularly estimated by the quasi-maximum likelihood method (QMLE) under the assumption that the distribution of one observation conditional to the past is normal. The asymptotic properties of the QMLE are well established. Weiss (1989) showed that QMLE estimates are consistent and asymptotically normal under fourth moment

Daniel Eni is a Senior Lecturer in the Department of Mathematics and Computer Science. Email him at: daneni58@yahoo.com.

conditions. These were again shown by Ling and McAleer (2003) under second moment conditions. If the assumption of normality is satisfied by the data, then the method will produce efficient estimates; otherwise, inefficient estimates will be produced. Engle and Gonzalez-Rivera (1991) studied the loss of estimation efficiency inherent in QMLE and concluded it may be severe if the distribution density is heavy tailed.

The QMLE estimator requires the use of a numerical optimization procedure which depends on different optimization techniques for implementation. This potentially leads to different estimates, as shown by Brooks, Burke and Perssand (2001) and McCullough and Renfro (1999). Both studies reported different QMLE estimates across various packages using different optimization routines. These techniques estimate time-varying variances in different ways and may result in different interpretations and predictions with varying implications to the economy. To resolve these problems, Eni and Etuk (2006) developed an Autocovariance Base Estimator (ABE) for estimating the parameters of GARCH models through an ARMA transformation of the GARCH model equation. The purpose of this article is to rate the performance of the ABE when the normality assumption is violated.

The Autocovariance Base Estimator (ABE)

Consider the GARCH (p, q) equation

$$h_t = w_0 + \sum_{i=1}^{p} \alpha_i e_{t-1}^2 + \sum_{j=1}^{q} \mathrm{B}_j h_{t-1}, \qquad (1)$$

or its ARMA (Max (p, q), q) transform

$$\varepsilon_t^2 = w_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + B_i)\varepsilon_{t-1}^2 - \sum_{j=1}^{q} B_J a_{t-1} + B_0 a_t \qquad (2)$$

$\epsilon \sim N(0, \sigma^2)$.

To obtain the autoregressive parameters, consider that the variance, $Var(\varepsilon_t^2 \varepsilon_{t-i}^2)$ for i > q in equation (2) will not contain the moving average parameter B_i. Hence, i = q + 1 … q + p is used to obtain the estimator:

$$\begin{bmatrix} V_{q+1} \\ V_{q+2} \\ \vdots \\ V_{q+\max(p,q)} \end{bmatrix} = \begin{bmatrix} Vq & V_{q-1} & \cdots & V_{q-(p-1)} \\ V_{q-1} & V_q & \cdots & V_{q-(p-2)} \\ \vdots & \vdots & \vdots & \\ V_{q-(p-1)} & V_{q-(p-2)} & \cdots & Vq \end{bmatrix} \begin{bmatrix} (\alpha_1 + \mathrm{B}_1) \\ (\alpha_2 + \mathrm{B}_2) \\ \vdots \\ (\alpha_p + \mathrm{B}q) \end{bmatrix} \qquad (3)$$

Where V_i is the set of variances associated with equation (2). The autoregressive parameters $\alpha_i + \mathrm{B}_i$ are obtained by solving (3).

Eni and Etuk (2006) have shown that the moving average parameters B_i can be obtained from

$$\sum_{i=0}^{p} f(\Phi_i) \begin{vmatrix} V_i & V_{i-1} & \cdots & V_{I-p} \\ V_{i+1} & V_i & \cdots & V_{i-(p-1)} \\ \vdots & \vdots & \vdots & \vdots \\ V_{i+q} & V_{i-q-1} & \cdots & V_{i+q-p} \end{vmatrix} \begin{bmatrix} \Phi_0 \\ -\Phi_1 \\ \vdots \\ -\Phi_p \end{bmatrix} =$$

$$\sigma_a^2 \begin{bmatrix} B_0 & -B_1 & \cdots & -B_{q-1} & -B_q \\ -B_1 & -B_2 & \cdots & -B_q & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -B_{q-1} & -B_q & \cdots & 0 & 0 \\ -B_q & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} B_0 \\ -B_1 \\ \vdots \\ -B_{q-1} \\ -B_q \end{bmatrix}$$

or

$$\sum_{i=0}^{p} f(\Phi_i) V\Phi = \sigma_a^2 B\, b \qquad (4)$$

where

$$f(\Phi_i) = -\Phi_i, f(\Phi_0) = \Phi_0, \Phi_i = (\alpha_i + \mathrm{B}_i).$$

Note that the quantity $\sum_{i=0}^{p} f(\Phi_i) V\Phi$ is known, the variance V having been calculated from the data, and the autoregressive parameters Φ having been calculated from equation (3).

The moving average parameters Bi are found by solving the system:

$$F(B)=\sum_{i=0}^{p} f\left(\Phi_i\right)V\Phi-\sigma_r^2 B\, b=0 \qquad (5)$$

Equation (5) is nonlinear and the solution can be found only through an iterative method. One procedure to consider is based on the Newton-Raphson algorithm, in this case, the B_{r+1} solution is obtained from the r^{th} approximation according to

$$B_{r+1}=B_r-\left\{f'(B_r)\right\}^{-1} f(B_r) \qquad (7)$$

where $f(B_r)$ and $f'(B_r)$ represent the vector function (5) and its derivative evaluated at $B=B_r$. Note that

$$f'(B)=\sigma_a^2\begin{bmatrix} B_0 & B_1 & \cdots & B_{q-1} & B_q \\ -B_1 & B_2 & \cdots & B_q & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ -B_{q-1} & B_q & \cdots & 0 & 0 \\ -B_q & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$+\sigma_a^2\begin{bmatrix} B_0 & B_1 & \cdots & B_{q-1} & B_q \\ 0 & -B_0 & \cdots & B_{q-z} & B_{q-1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & -B_0 & B_1 \\ 0 & 0 & \cdots & 0 & -B_0 \end{bmatrix}$$

$$=\sigma_r^2\,(D_1+D_2)$$

and (4) becomes

$$B_{r+1}=B_r+\left\{\sigma_a^2(D_1+D_2)\right\}^{-1}\left[\sum_{i=0}^{p} f(\Phi_1)V\,\Phi-\sigma_a^2 B\, b\right] \qquad (8)$$

The starting point for the iteration (8) is $\sigma_r^2=1$, $B_0=V_0$, $B_i=0$, $i=1\ldots q$.

Having computed the Autoregressive parameters $\Phi_i=(\alpha_i+B_i)$ and the Moving average parameter B_i, it is simple to obtain the GARCH (p, q) parameters, α_i, and the constant parameter w_0, which is estimated using

$$W=E(\varepsilon_t^2)(1-\sum_{i=1}^{p}\alpha_i-\sum_{i=1}^{B}B_i). \qquad (9)$$

Methodology

The data generating process (DGP) in this study involved the simulation of 1,500 data points with 10 replications using the random number generator in MATLAB 5. The random number generator in MATLAB 5 is able to generate all floating point numbers in the interval $\left[2^{-53},\ 1-2^{-53}\right]$. Thus, it can generate 2^{1492} values before repeating itself. Note that 1,500 data points are equivalent to $2^{10.55}$ and, with 10 replications; results in only $2^{13.87267}$ data points. Hence 1,500 data points with 10 replications were obtained without repetitions. Also, a program implementation was used for ARMA to find the QMLE (McLeod and Sales, 1983). Although normality would typically be assumed, the data points were simulated using the Log normal and the T-distribution with 5, 10 and 15 degrees of freedom.

Of the 1,500 data points generated for each process, the first 200 observations were discarded to avoid initialization effects, yielding a sample size of 12,000 observations, with results reported in sample sizes of 200, 500, 1,000 and 1,200. These sample presentations enable tracking of consistency and efficiency of the estimators. The relative efficiency of the autocovariances based estimator (ABE) and the quasi-maximum likelihood (QML) estimators were studied under this misspecification of

distribution function. The selection criteria used was the Aikake information criteria (AIC).

For simulating the data points, the conditional variance equation for low persistence due to Engle and Ng (1993) was adopted.

$$h_t = 0.2 + 0.05\,\varepsilon_{t-1}^2 + 0.75 h_{t-1}$$

$$\varepsilon_{t-1}^2 = h_t Z_t^2$$

and Z_t^2 is any of $Z \sim t_5$ or Z~t_{10} or Z~LN(0,1) or Z~t_{15}, where N = normality, t_V = t-distribution with V degree of freedom, and LN = log normal.

Results

Apart from the parameter setting in the DGP, selected studies of the parameter settings $(W, \alpha, B) = (0,\ 1,\ 0.15,\ 0.85)$ and $(W, \alpha, B) = (0,\ 1,\ 0.25,\ 0.65)$ (Lumsdaine, 1995), and $(W, \alpha, B) = (1, 0.3, 0.6)$ and $(W, \alpha, B) = (1, 0.05, 0.9)$ (Chen, 2002) were also studied. The results obtained agree with the results obtained from detailed studies of the DGP.

Table 1 shows the results from a sample size of 200 data points. The table reveals that the estimates are poor for QMLE and ABE. On the basis of the Aikate information criteria (AIC), however, the QMLE performed better than the ABE except under the log normal distribution where ABE performed better than QMLE.

Table 2 shows that the estimates using a sample size of 500 are better, although still poor. The performance bridge between QMLE and ABE appears to be closing. This is observed from the AIC of QMLE and ABE under the different probability distribution functions, with one exception in the case of the log normality. Surprisingly, the QMLE method failed to show consistency, but it is notable that the performance of both methods was enhanced under the t-distribution as the degrees of freedom increase.

Table 3 shows that both estimation models, QMLE and ABE, had equal performance ratings and gave consistent estimates in general. However, the ABE had an edge in its performance under $t_{(5)}$ and LN(0, 1) while QMLE had an edge under $t_{(10)}$ and $t_{(15)}$. The estimates under $t_{(15)}$ and $t_{(10)}$ were close to their true values for both estimation methods. Finally, the results shown in Table 3 are further confirmed by examining Table 4 where the two methods have nearly equal ratings based on the values of their AIC.

Conclusion

It is shown is this study that the ABE method is adequate in estimating GARCH model parameters and can perform as well as the maximum likelihood estimate for reasonably large numbers of data points when the distribution assumption is misspecified.

Table 1: Performance Rating of QMLE and ABE for Sample Size *n* = 200

	Estimation Method							
	QMLE				ABE			
Estimates	W	α	B	AIC	W	α	B	AIC
t (5)	0.16	0.01	0.77	-70.90	1.14	0.016	0.74	-65.312
t (10)	0.14	0.014	0.76	-140.36	1.138	0.012	0.75	-124.31
t (15)	0.15	0.17	0.76	-169.40	1.42	0.016	0.76	-157.21
Ln (0, 1)	9.3	-0.2	0.86	129.17	6.2	0.20	0.81	108.23

Table 2: Performance Rating of QMLE and ABE for Sample Size n = 500

	Estimation Method							
	QMLE				ABE			
Estimates	W	α	B	AIC	W	α	B	AIC
t (5)	0.115	0.02	0.739	-132.341	0.15	0.025	0.73	-151.24
t (10)	0.018	0.034	0.742	-1021.22	0.21	0.029	0.74	-956.31
t (15)	0.17	0.036	0.75	-1973.42	0.20	0.030	0.75	-1472.40
Ln (0, 1)	6.79	-0.15	0.88	289.39	3.94	0.08	0.80	108.21

Table 3: Performance Rating of QMLE and ABE for Sample Size n = 1,000

	Estimation Method							
	QMLE				ABE			
Estimates	W	α	B	AIC	W	α	B	AIC
t (5)	0.18	0.02	0.74	-137.12	0.19	0.03	0.73	-140.12
t (10)	0.193	0.029	0.75	-1141.62	0.22	0.03	0.74	-1094.72
t (15)	0.195	0.034	0.75	-1984.71	0.21	0.04	0.75	-1976.22
Ln (0, 1)	5.24	-0.40	0.86	119.72	3.50	0.07	0.79	101.13

Table 4: Performance Rating of QMLE and ABE for Sample Size n = 1,200

	Estimation Method							
	QMLE				ABE			
Estimates	W	α	B	AIC	W	α	B	AIC
t (5)	0.15	0.03	0.76	-162.11	0.18	0.039	0.73	-173.70
t (10)	0.018	0.039	0.743	-1391.30	0.19	0.041	0.74	-1350.11
t (15)	0.19	0.043	0.746	-2441.30	0.19	0.044	0.742	-2430.39
Ln (0, 1)	4.3	0.08	0.81	168.59	3.48	0.060	0.78	256.23

References

Baillie, R., Bollerslev, T., & Mikkelson, H. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*, 3-30.

Bollerslev, T. (1986). Generalized autoregressive heteroskedasticity. *Journal of Econometrics*, *31*, 307-327.

Brooks, I., Burke, S., & Perssand, G. (2001). Benchmarks and accuracy of GARCH model estimation. *International Journal of Forecasting*, *17*, 45-56.

Chen, Y. T. (2002). On the robustness of Ljung-Box and Mcleed-Li Q test. *Economics Bulletin*, *3*(*17*), 1-10.

Engle, F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, *50*, 987-1008.

Engle, F., & Gonzalez-Revera. (1991). Semi parametric ARCH models. *Journal of Business and Economic Statistics*, *19*, 3-29.

Engle, F., & Ng, V. (1993). Measurement and testing the impact of news on volatility. *Journal of Finance*, *48*(*5*), 1749-1778.

Gosten, L., Jaganatan, R., & Runkle, D. (1993). On the relationship between the expected value and the volatility of the nominal excess return on stock. *Journal of Finance*, *48*, 1779-1801.

Eni, D., & Etuk, E. H. (2006). *Parameter estimation of GARCH models: An autocovariance approach*. Proceedings of International Conference on New Trends in the Mathematical & Computer Sciences with Applications to Real World Problems Held at Covenant University Ota, Nigeria, 357-368.

Ling & McAleer, M. (2003). Asymptotic theory for a vector GARCH (1.1) quasi-maximum likelihood estimator. *Economic Theory*, *19*, 280-310.

Lumsdain, R. L. (1995). Consistency and asymptotic normality of the quasi-maximum likelihood estimator. *Econometrica*, *64*, 575-596.

McCullough, B. D., & Renfro, C. G. (1999). Bench marks and software standard: A case study of GARCH procedure. *Journal of Economics and Social Measurement*, *25*, 59-71.

McLeod, A., & Sales, P. (1983). Algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Journal of Applied Statistics*, *32*, 211-2190

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*, 347-370

Weiss, A. (1986). Asymptotic Theory of GARCH (1,1) model. *Economic Theory*, *2*, 107-131

Zakoian, J. M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamic and Control*, *18*, 931-955.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 566-570

1538 – 9472/09/$95.00

Test for the Equality of the Number of Signals

Madhusudan Bhandary
Columbus State University

Debasis Kundu
Indian Institute of Technology
Kanpur, India

A likelihood ratio test for testing the equality of the ranks of two non-negative definite covariance matrices arising in the area of signal processing is derived. The asymptotic distribution of the test statistic follows a Chi-square distribution from the general theory of likelihood ratio test.

Key words: Likelihood ratio test; signal processing; white noise.

Introduction

In the area of signal processing, signals are observed at different sensors from different sources at different time points. Wax, Shan and Kailath (1984) and Whalen (1971) discussed models and varieties of problems in signal processing. In general, the signal processing model is as follows:

$$\mathbf{X}(t) = A\mathbf{S}(t) + \mathbf{n}(t) \qquad (1)$$

where, $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_p(t))'$ is the px1 observation vector at time t, $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_q(t))'$ is the qx1 vector of unknown random signals at time t, $\mathbf{n}(t) = (n_1(t), n_2(t), \ldots, n_p(t))'$ is the px1 random noise vector at time t, $A = (A(\Phi_1), A(\Phi_2), \ldots, A(\Phi_q))$ is the pxq matrix of unknown coefficients, and $A(\Phi_r)$ is the px1 vector of functions of the elements of unknown vector Φ_r associated with the r^{th} signal and $q < p$.

In model (1), $\mathbf{X}(t)$ is assumed to be distributed as p-variate normal with mean vector zero and dispersion matrix $A\Psi A' + \sigma^2 I_p = \Gamma + \sigma^2 I_p$, where $\Gamma = A\Psi A'$ is unknown n.n.d. matrix of rank q(<p) and Ψ = covariance matrix of $\mathbf{S}(t)$, σ^2 (>0) is unknown and $\sigma^2 I_p$ is the covariance matrix of the noise vector $\mathbf{n}(t)$. In this case, the model is called white noise model.

One important problem that arises in the area of signal processing is to estimate q, the number of signals transmitted. This problem is equivalent to estimating the multiplicity of the smallest eigenvalue of the covariance matrix of the observation vector. Anderson (1963), Krishnaiah (1976) and Rao (1983), among others, considered this problem. Wax and Kailath (1984) and Zhao, et al. (1986a, b) used information theoretic criteria proposed by Akaike (1972), Rissanen (1978) and Schwartz (1978) to estimate the number of signals.

More recently, Chen, et al. (2001), Chen (2002) and Kundu (2000) developed procedures for estimating the number of signals. This article considers the two sample problem of testing the equality of the number of signals between two sets of data from two populations. This problem is relevant in practice in the area of signal processing because it is important to know whether the total numbers of signals received are the same or not for two different days, which can be separated by a lengthy time. This problem is equivalent to testing the equality of multiplicity of the smallest eigenvalue of the covariance matrices of observation vectors of the two sets of data. Consider the following model:

$$X_i(t) = A_i S_i(t) + N_i(t); i = 1, 2$$

Madhusudan Bhandary is an Associate Professor in the Department of Mathematics at Columbus State University, Columbus, GA 31907. Email: bhandary_madhusudan@colstate.edu. Debasis Kundu is a Professor in the Department of Mathematics. E-mail: kundu@iitk.ac.in.

where, $X_i(t)$ is a px1 observation vector for the i[th] population, $A_i = (A_i(\Phi_1^i),...,A_i(\Phi_{q_i}^i))$, $S_i(t) = (S_1^i(t),...,S_{q_i}^i(t))$, $i = 1,2$ and $N_i(t)$ is a px1 random noise vector for the i[th] population. The following are assumed about $N_i(t)$ and $S_i(t)$:

$$\mathrm{N_i}(t) \sim N_p(0,\Sigma_i),$$
$$S_i(t) \sim N_p(0,\Psi_i),$$

and $N_i(t)$ and $S_i(t)$ are independently distributed. The null hypothesis to test is $H_{0k}: q_1 = q_2 = k$, and the alternative hypothesis is $H_1: q_1 \neq q_2, k = 0,1,...,p-1$. At this point, the likelihood ratio test is derived next and asymptotic distribution of the test statistic is used to obtain the critical value.

Likelihood Ratio Test: Case 1

Consider $\Sigma_i = \sigma^2 I_p, i = 1,2$. The test hypotheses are:

$$H_{0k}: q_1 = q_2 = k$$

and

$$H_1: q_1 \neq q_2, k = 0,1,...,p-1.$$

The observations from the two populations are as follows: $X_1(t_1),...,X_1(t_{n_1})$ are i.i.d. $\sim N_p(0, A_1\Psi_1A_1' + \sigma^2 I_p)$ and $X_2(t_{n_1+1}),...,X_2(t_{n_1+n_2})$ are i.i.d. $\sim N_p(0, A_2\Psi_2A_2' + \sigma^2 I_p)$.

Let $R_i = A_i\Psi_i A_i' + \sigma^2 I_p, i = 1,2$. It may be stated that testing H_{0k} is equivalent to testing the rank of $A_i\Psi_i A_i'$ = k, i = 1, 2.

Let $R_i = R_i^{(k)}$ under H_{0k}, i = 1,2. Using spectral decomposition of $R_1^{(k)}$ and $R_2^{(k)}$, it can be written that

$$R_1^{(k)} = (\lambda_1 - \sigma^2)U_1U_1' + ... + (\lambda_k - \sigma^2)U_kU_k' + \sigma^2 I_p$$

and

$$R_2^{(k)} = (\mu_1 - \sigma^2)V_1V_1' + ... + (\mu_k - \sigma^2)V_kV_k' + \sigma^2 I_p$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k > \sigma^2$ are the eigenvalues of $R_1^{(k)}$ and $U_1,...,U_k$ are the corresponding orthonormal eigenvectors of $A_1\Psi_1A_1'$ and similarly, $\mu_1 \geq \mu_2 \geq ... \geq \mu_k > \sigma^2$ are the eigenvalues of $R_2^{(k)}$ and $V_1,...,V_k$ are the corresponding orthonormal eigen vectors of $A_2\Psi_2A_2'$.

Thus, under H_{0k} the log-likelihood (apart from a constant term) is

$$\begin{aligned}\log L &= -\frac{n_1}{2}\log\left|R_1^{(k)}\right| - \frac{n_1}{2}tr.(\hat{R}_1(R_1^{(k)})^{-1}) \\ &\quad -\frac{n_2}{2}\log\left|R_2^{(k)}\right| - \frac{n_2}{2}tr.(\hat{R}_2(R_2^{(k)})^{-1}) \\ &= -\frac{n_1}{2}tr.(\hat{R}_1(R_1^{(k)})^{-1}) - \frac{n_2}{2}tr.(\hat{R}_2(R_2^{(k)})^{-1}) \\ &\quad -\frac{n_1}{2}\sum_{i=1}^{k}\log\lambda_i - \frac{n_2}{2}\sum_{i=1}^{k}\log\mu_i \\ &\quad -\frac{n}{2}(p-k)\log\sigma^2 \end{aligned} \tag{2}$$

where

$$\hat{R}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}X_1(t_i)X_1'(t_i),$$

$$\hat{R}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}X_2(t_i)X_2'(t_i)$$

and

$$n = n_1 + n_2.$$

Rather than maximizing (2), equivalently minimize

$$\begin{aligned}\log L^* &= n_1tr.(\hat{R}_1(R_1^{(k)})^{-1}) + n_2tr.(\hat{R}_2(R_2^{(k)})^{-1}) \\ &\quad + n_1\sum_{i=1}^{k}\log\lambda_i + n_2\sum_{i=1}^{k}\log\mu_i \\ &\quad + n(p-k)\log\sigma^2 \end{aligned} \tag{3}$$

Orthogonal matrices P_1, P_2, U_1, U_2 exist such that

$$\hat{R}_1 = P_1 D_1 P_1', \qquad R_1^{(k)} = U_1 \Lambda_1 U_1'$$

and

$$\hat{R}_2 = P_2 D_2 P_2', \qquad R_2^{(k)} = U_2 \Lambda_2 U_2'$$

where, $D_1 = Diag.(l_1,...,l_p), \quad l_1 \geq l_2 \geq ... \geq l_p$ are the eigenvalues of $\hat{R}_1$ and similarly, $D_2 = Diag.(\xi_1,...,\xi_p), \quad \xi_1 \geq \xi_2 \geq ... \geq \xi_p$ are the eigenvalues of $\hat{R}_2$ $\Lambda_1 = Diag.(\lambda_1,...,\lambda_k,\sigma^2,...,\sigma^2)$, and $\Lambda_2 = Diag.(\mu_1,...,\mu_k,\sigma^2,...,\sigma^2)$. Thus, (3) can be rewritten as follows:

$$\begin{aligned}\log L^* &= n_1 tr.(P_1 D_1 P_1' U_1 \Lambda_1^{-1} U_1') \\ &\quad + n_2 tr.(P_2 D_2 P_2' U_2 \Lambda_2^{-1} U_2') \\ &\quad + n_1 \sum_{i=1}^{k} \log \lambda_i + n_2 \sum_{i=1}^{k} \log \mu_i \\ &\quad + n(p-k) \log \sigma^2 \\ &= n_1 tr.(D_1 V_1 \Lambda_1^{-1} V_1') + n_2 tr.(D_2 V_2 \Lambda_2^{-1} V_2') \\ &\quad + \text{term independent of } V_1 \text{and} V_2 \end{aligned} \tag{4}$$

where, $V_1 = P_1' U_1$ and $V_2 = P_2' U_2$ and hence V_1 and V_2 are orthogonal.

Differentiating (4) with respect to V_1 subject to $V_1 V_1' = I_p$ and equating it to 0, results in

$$D_1 \left[\Lambda_1^{-1} V_1' - V_1 \Lambda_1^{-1} V_1^{-2} \right] = 0$$

$$i.e., V_1 = I_p$$

Similarly, $V_2 = I_p$ is obtained. Hence, given $\lambda_i's, \mu_i's$ and σ^2,

$$\begin{aligned}&\underset{H_{0k}}{Inf} \log L^* = \\ &n_1 \left(\sum_{i=1}^{k} \frac{l_i}{\lambda_i} + \frac{\sum_{i=k+1}^{p} l_i}{\sigma^2} \right) + n_2 \left(\sum_{i=1}^{k} \frac{\xi_i}{\mu_i} + \frac{\sum_{i=k+1}^{p} \xi_i}{\sigma^2} \right) \\ &+ n_1 \sum_{i=1}^{k} \log \lambda_i + n_2 \sum_{i=1}^{k} \log \mu_i + n(p-k) \log \sigma^2 \end{aligned} \tag{5}$$

Differentiating (5) with respect to $\lambda_i's$ and equating it to 0, results in

$$-n_1 \frac{l_i}{\lambda_i^2} + \frac{n_1}{\lambda_i} = 0$$

$$i.e., \hat{\lambda}_i = l_i$$

Similarly, $\hat{\mu}_i = \xi_i; i = 1,...,k.$

Differentiating (5) with respect to σ^2 and equating it to 0, results in

$$\hat{\sigma}^2 = \frac{n_1 \sum_{i=k+1}^{p} l_i + n_2 \sum_{i=k+1}^{p} \xi_i}{n(p-k)},$$

hence,

$$\begin{aligned}\underset{H_{0k}}{Sup} \log L &= -nk - n_1 \frac{\sum_{i=k+1}^{p} l_i}{\sigma^2} - n_2 \frac{\sum_{i=k+1}^{p} \xi_i}{\sigma^2} \\ &\quad - n_1 \sum_{i=1}^{k} \log l_i - n_2 \sum_{i=1}^{k} \log \xi_i - n(p-k) \log \hat{\sigma}^2 \\ &= -np - n_1 \sum_{i=1}^{k} \log l_i - n_2 \sum_{i=1}^{k} \log \xi_i \\ &\quad - n(p-k) \log \left(\frac{n_1 \sum_{i=k+1}^{p} l_i + n_2 \sum_{i=k+1}^{p} \xi_i}{n(p-k)} \right) \\ &= L_1 \text{ (say)} \end{aligned}$$

In the above expression of L_1, the unknown k can be estimated by using Zhao, Krishnaiah and Bai's (1986a,b) information criterion as follows: Estimate k by $\hat{k}$ such that

$$I(\hat{k}, c_n) = \min_{0 \le k \le p-1} I(k, c_n),$$

where

$$I(k, c_n) = -L_1 + c_n \left\{ (2k+1) + 2\left(pk - k - \frac{k(k-1)}{2} \right) \right\}$$

and c_n is such that

$$\text{(i)} \lim_{n \to \infty} \frac{c_n}{n} = 0$$

$$\text{(ii)} \lim_{n \to \infty} \frac{c_n}{\log \log n} = \infty$$

For practical purposes, choose $c_n = \log n$ which satisfies conditions (i) and (ii). Hence,

$$L_1^* = -np - n_1 \sum_{i=1}^{\hat{k}} \log l_i - n_2 \sum_{i=1}^{\hat{k}} \log \xi_i - n(p-\hat{k}) \log \left(\frac{n_1 \sum_{i=\hat{k}+1}^{p} l_i + n_2 \sum_{i=\hat{k}+1}^{p} \xi_i}{n(p-\hat{k})} \right) \quad (6)$$

Similarly,

$$L_2^* = \sup_{H_1} \log L = -np - n_1 \sum_{i=1}^{\hat{q}_1} \log l_i - n_2 \sum_{i=1}^{\hat{q}_2} \log \xi_i - \left[n_1 (p-\hat{q}_1) + n_2 (p-\hat{q}_2) \right] \log \left(\frac{n_1 \sum_{i=\hat{q}_1+1}^{p} l_i + n_2 \sum_{i=\hat{q}_2+1}^{p} \xi_i}{n_1 (p-\hat{q}_1) + n_2 (p-\hat{q}_2)} \right) \quad (7)$$

where, $\hat{q}_1$ and $\hat{q}_2$ are obtained such that

$$I(\hat{q}_1, \hat{q}_2, c_n) = \min_{\substack{0 \le q_1 \le p-1 \\ 0 \le q_2 \le p-1}} I(q_1, q_2, c_n)$$

and

$$I(q_1, q_2, c_n) = np + n_1 \sum_{i=1}^{q_1} \log l_i + n_2 \sum_{i=1}^{q_2} \log \xi_i + \begin{bmatrix} n_1(p-q_1) \\ +n_2(p-q_2) \end{bmatrix} \log \left(\frac{n_1 \sum_{i=q_1+1}^{p} l_i + n_2 \sum_{i=q_2+1}^{p} \xi_i}{n_1(p-q_1) + n_2(p-q_2)} \right) + c_n \begin{Bmatrix} (q_1+q_2)+1 \\ +(p-1)(q_1+q_2) - \frac{q_1(q_1-1)}{2} - \frac{q_2(q_2-1)}{2} \end{Bmatrix}$$

and c_n is defined the same as previously.

Hence log of likelihood ratio statistic is $L_1^* - L_2^*$, where L_1^* and L_2^* are given by (6) and (7) respectively. The critical value for this test can be approximated from the fact that asymptotically, $-2(L_1^* - L_2^*) \sim \chi^2_{\gamma(\hat{q}_1, \hat{q}_2, \hat{k})}$ under H_0 where,

$$\gamma(\hat{q}_1, \hat{q}_2, \hat{k}) = (\hat{q}_1 + \hat{q}_2 - 2\hat{k}) + (p-1)(\hat{q}_1 + \hat{q}_2 - 2\hat{k}) + \hat{k}(\hat{k}-1) - \frac{\hat{q}_1(\hat{q}_1-1)}{2} - \frac{\hat{q}_2(\hat{q}_2-1)}{2}.$$

Likelihood Ratio Test: Case 2

Consider, $\Sigma_i = \sigma_i^2 I_p, i = 1, 2$. For case 2, the problem can be solved similarly and the problem is easier than that in case 1.

Likelihood Ratio Test: Case 3

Consider, σ^2 is known in case 1. Without loss of generality, σ^2 = 1 can be assumed and in that case, the log likelihood must be maximized with respect to the eigenvalues subject to the condition that the eigenvalues are greater than 1, in which case the technique presented by Zhao, Krishnaiah and Bai (1986a, b) can be used.

References

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the second international symposium on information theory, supp. to problems of control and information theory*, 267 – 281.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, *34*, 122-138.

Chen, P. (2002). A selection procedure for estimating the number of signal components. *Journal of Statistical Planning and Inference*, *105*, 299-301.

Chen, P., Wicks, M. C., & Adve, R. S. (2001). Development of a statistical procedure for detecting the number of signals in a radar measurement. *IEEE Proceedings of Radar, Sonar and Navigations*, *148*(*4*), 219-226.

Krishnaiah, P. R. (1976). Some recent developments on complex multivariate distributions. *Journal of Multivariate Analysis*, *6*, 1-30.

Kundu, D. (2000). Estimating the number of signals in the presence of white noise. *Journal of Statistical Planning and Inference*, *90*, 57-68.

Rao, C.R. (1983). Likelihood ratio tests for relationships between two covariance matrices. In T. Amemiya, S. Karlin & L. Goodman (Eds.) *Studies in Econometrics*, *Time Series and Multivariate Statistics.* New York: Academic Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 463-471.

Schwatrz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Wax, M., & Kailath, T. (1985). Determination of the number of signals by information theoretic criteria. *IEEE Trans. Acoustics Speech Signal Processing*, *33*, 387-392.

Wax, M., Shan, T. J., & Kailath, T. (1984). Spatio temporal spectral analysis by eigen structure methods. *IEEE Trans. Acoustics Speech Signal Processing*, *32*, 817-827.

Whalen, A. D. (1971). Detection of signals in noise. New York: Academic Press.

Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986a). On detection of number of signals in presence of white noise. *Journal of Multivariate Analysis*, *20*, 1-25.

Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986b). On detection of the number of signals when the noise covariance matrix is arbitrary, *Journal of Multivariate Analysis*, *20*, 26-49.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 571-582

1538 – 9472/09/$95.00

A Linear *B*-Spline Threshold Dose-Response Model with Dose-Specific Response Variation Applied to Developmental Toxicity Studies

Chin-Shang Li
University of California, Davis

Daniel L. Hunt
Radiation Therapy Oncology Group, Philadelphia, PA

A linear *B*-spline function was modified to model dose-specific response variation in developmental toxicity studies. In this new model, response variation is assumed to differ across dose groups. The model was applied to a developmental toxicity study and proved to be significant over the previous model of singular response variation.

Key words: Developmental toxicity study, dose-response, interior knot, linear *B*-spline, response variation, threshold.

Introduction

In a developmental toxicity study, fetal response is measured and recorded in each litter of an animal that has been directly exposed to some toxic substance that is environmentally ambient and that poses a developmental threat. Developmental endpoints include death, abnormality (all types), weight and length. A positive fetal response, equivalent to negative indicators of these endpoints, implies negative reaction to the toxic substance. Upon study execution, the fetal risk of indirect exposure can be assessed. The U.S. Environmental Protection Agency (USEPA) uses such study results to determine safety exposure levels for the general population (USEPA, 1991); statistical modeling is a key factor in estimating risk (Ryan, 2000).

The default assumption in the risk assessment process for developmental toxicity studies is that a threshold dose level exists (USEPA, 1991). Threshold is the maximum dose level at which the response is equivalent to the background response. The USEPA uses the no-observed-adverse-effects-level (NOAEL) and benchmark dosing approaches. The NOAEL approach identifies the highest dose level at which the response is not statistically significant from the control. Benchmark dosing employs actual dose-response modeling. First proposed by Crump (1984), the benchmark dose is a lower confidence limit for the dose equivalent to a level that yields an acceptable limit excess risk. Although both approaches search for a tolerable dose level, neither is a pure threshold model.

Cox (1987) introduced a variety of pure threshold models for application to toxicology studies. Schwartz, et al. (1995) applied a threshold model to a developmental toxicity study, using quasi-likelihood techniques for estimating model parameters. Hunt and Rai (2003) introduced the threshold dose-response model with a single parameter for response variation included in the dose-response function. All these approaches model the behavior of the dose-response pattern below the threshold level as one of constant response. The model proposed in this study inherently estimates the threshold, while tracking the change in the slope of the dose-response curve, thereby allowing more flexibility in the sense of being able to model multiple dose-response shapes.

Chin-Shang Li is an Associate Professor in the Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, CA. Email: cssli@ucdavis.edu. Daniel L. Hunt is a Senior Statistician in the RTOG Statistical Center, American College of Radiology, Philadelphia, PA. Email: dhunt@acr.org.

Hunt and Rai (2003) first modeled the observed variation by including a parameter that equated to the interlitter response variation. Subsequently, they modified their model to include additional parameters to account for the noticeable multiple response variation across dose groups (Hunt & Rai, 2007). They reported significance with the model of dose-specific response variation compared to a model with uniform variation across dose levels. They also conducted simulations and found that the dose-specific model with multiple variation parameters led to unbiased estimation of all model parameters, whether the true variance structure was that of single or multiple parameters, whereas the single-parameter model was less robust. These results are similar to those of Kupper, et al. (1986), who assumed the beta-binomial distribution for the response number for each litter and a logistic dose-response model and found that the model with the multiple-intralitter correlation structure produced less biased results than the model that assumed a single-intralitter correlation.

Similar to Hunt and Rai (2007), the model used in this study includes dose-specific parameters that estimate the response variation in addition to using polynomial regression splines to fit to the dose-response pattern (Ramsay, 1988). The regression spline approach was used formerly in a model with one parameter for response variation (Li & Hunt, 2004). As the polynomial, degree one (linear) was used and a set of *B*-splines was constructed recursively to help fit the model. The theory of *B*-splines is described in de Boor (2001). Integral to this theory is the incorporation of interior knots as change points in the direction of the plotted curve. The ability to incorporate these knots is desirable for data from developmental studies as the threshold is inherently assumed.

Methodology

In a developmental toxicity study, there are g dose groups, each of which has a certain level of a toxic substance. The i^{th} dose group contains m_i animals, and therefore litters (i = 1, …, g). For n_{ij} implantations of the j^{th} animal (j = 1, …, m_i) in the i^{th} dose group, let x_{ij} be the number of fetuses that experience at least one adverse effect. Adverse effects include early and late fetal death and any kind of malformation (morphological, visceral or skeletal). An adverse effect such as death supersedes malformation. If $P_j(d_i)$ be the probability of a fetus in the j^{th} litter indirectly exposed to the i^{th} dose level, d_i, experiencing an adverse event, then the proposed dose-response model is the following:

$$P_j(d_i) = \frac{\exp(\sum_{k=1}^{3}\theta_k B_{k,2}(d_i,\xi)+\sigma_i z_{ij})}{1+\exp(\sum_{k=1}^{3}\theta_k B_{k,2}(d_i,\xi)+\sigma_i z_{ij})} = \frac{\exp(\mathbf{B}_2(d_i,\xi)\boldsymbol{\theta}+\sigma_i z_{ij})}{1+\exp(\mathbf{B}_2(d_i,\xi)\boldsymbol{\theta}+\sigma_i z_{ij})} \quad (1)$$

Here,

$$\mathbf{B}_2(d_i,\boldsymbol{\xi}) = \left(B_{1,2}(d_i,\boldsymbol{\xi}), B_{2,2}(d_i,\boldsymbol{\xi}), B_{3,2}(d_i,\boldsymbol{\xi})\right)$$

is the set of (order 2, degree 1) linear *B*-splines, with 1 interior knot, ξ, defined on the dose interval $[d_1, d_g)$, and derived recursively from the order 1 (degree 0) *B*-splines $B_{k,1}(d_i,\xi)$. If $\xi_1 = \xi_2 = d_1$, $\xi_3 = \xi$, and $\xi_4 = \xi_5 = d_g$, then, the order 1 *B*-splines are given by:

$$B_{k,1}(d_i,\xi) = \begin{cases} 1, & d_i \in [\xi_k, \xi_{k+1}) \\ 0, & \text{otherwise} \end{cases} \quad k = 1, 2, 3, \quad (2)$$

and the order 2 *B*-splines formed recursively from (2) are given by:

$$B_{k,2}(d_i,\xi) = \frac{d_i-\xi_k}{\xi_{k+1}-\xi_k} B_{k,1}(d_i,\xi) + \frac{\xi_{k+2}-d_i}{\xi_{k+2}-\xi_{k+1}} B_{k+1,1}(d_i,\xi) \quad (3)$$

Also from (1), the number of elements in the three-parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$, for the linear *B*-splines, corresponds to the degree, 1, plus the order, 2 (see de Boor, 2001). The interior knot ξ represents a change point in the direction of the dose-response relationship. The modification in equation (1) from the previous model is in the parameter(s), σ_i, coefficients of $z_{ij} \sim N(0,1)$. Thus, with the subscript *i*, the response variability σ_i^2 is allowed to differ across dose levels $i = 1, \ldots, g$.

The likelihood function based on the dose-response model in equation (1) is given by:

$$L(\boldsymbol{\theta}, \boldsymbol{\sigma}, \xi) = \prod_{i=1}^{g} \prod_{j=1}^{m_i} \binom{n_{ij}}{x_{ij}} \int_{-\infty}^{\infty} P_j^{x_{ij}}(d_i)[1 - P_j(d_i)]^{n_{ij} - x_{ij}} \frac{\exp(-z_{ij}^2 / 2)}{\sqrt{2\pi}} dz_{ij} \quad (4)$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_g)$. Also note, the interior knot ξ is regarded as a parameter of the model that must be estimated in addition to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$. The likelihood function in equation (4) integrates out the random effect z_{ij} from the joint distribution, thereby leaving a marginal function for the number of fetal responses x_{ij} (see Collett 1991, p. 208).

Because (4) cannot be solved directly, an approximation is used via the Gauss-Hermite formula for numeric integration, given by:

$$\int_{-\infty}^{\infty} f(u) \exp(-u^2) du = \sum_{l=1}^{q} a_l f(b_l). \quad (5)$$

Here q, the values of which a_l and b_l depend, is chosen to approximate (5). The standardized tables from which the values of q, a_l, and b_l may be found are in Abramowitz and Stegun (1972).

To approximate, first let $z_{ij} = u\sqrt{2}$, based on equation (5), take the log of the likelihood function in equation (4), and approximate the log-likelihood function by:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\sigma}, \xi) \approx -\frac{N}{2} \times \log \pi + \sum_{i=1}^{g} \sum_{j=1}^{m_i} \log \binom{n_{ij}}{x_{ij}} + \tilde{\ell}(\boldsymbol{\theta}, \boldsymbol{\sigma}, \xi), \quad (6)$$

where $N = \sum_{i=1}^{g} \sum_{j=1}^{m_i} n_{ij}$ the study sample size, and

$$\tilde{\ell}(\boldsymbol{\theta}, \boldsymbol{\sigma}, \xi) = \sum_{i=1}^{g} \sum_{j=1}^{m_i} \log \left(\sum_{l=1}^{q} a_l \frac{\left[\exp\left(\sum_{k=1}^{3} \theta_k B_{k,2}(d_i, \xi) + \sigma_i b_l \sqrt{2} \right) \right]^{x_{ij}}}{\left[1 + \exp\left(\sum_{k=1}^{3} \theta_k B_{k,2}(d_i, \xi) + \sigma_i b_l \sqrt{2} \right) \right]^{n_{ij}}} \right) \quad (7)$$

A value of q = 20 was chosen for the approximation; this value has been deemed acceptable in many settings (Collett, 1991).

A profile-likelihood approach was used to maximize $\tilde{\ell}(\boldsymbol{\theta}, \boldsymbol{\sigma}, \xi)$ in equation (7) and to estimate the parameters of the model. The approach begins with a search over the dose interval $(0, d_g)$, the domain of ξ. A large number of G grid points $\{\xi_t^* : t = 1, \ldots, G\}$ were chosen by using the formula: $\xi_t^* = d_g \times t / (G+1)$. Once a fixed grid point ξ_t^* was selected, the order 1 and 2 *B*-splines in equations (2) and (3) were calculated by regarding the fixed grid point ξ_t^* as the interior knot for that part of the search. As a result, the maximizer of the profile-likelihood $\tilde{\ell}(\boldsymbol{\theta}, \boldsymbol{\sigma}, \xi_t^*)$, denoted by $(\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\sigma}}_t)$, can be found and it yields $\tilde{\ell}(\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\sigma}}_t, \xi_t^*)$, $t = 1, 2, \ldots, G$. Hence, the maximum likelihood estimates are given by:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}, \hat{\xi}) = \arg\max_{\{(\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\sigma}}_t, \xi_t^*) : t = 1, 2, \ldots, G\}} \tilde{\ell}(\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\sigma}}_t, \xi_t^*). \quad (8)$$

To maximize the log-likelihood function in equation (6), the Olsson version (1974) of the Nelder-Mead simplex algorithm (1965) was used. The approach minimizes a function by

construction of a simplex of points; the dimension for each corresponds to the number of parameters that must be estimated. The magnitude of the simplex is the number of parameters + 1. Maximization of $\tilde{\ell}$ was accomplished by minimizing $-\tilde{\ell}$ using the simplex algorithm. This algorithm is coded in FORTRAN, version 90.

The asymptotic variance of the estimates in $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\sigma}},\hat{\xi})$ was obtained by evaluating $\mathbf{I}^{-1}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$, the inverse of the observed information matrix, at $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\sigma}},\hat{\xi})$. (The formula for $\mathbf{I}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ is in the Appendix.) The algorithm for computing asymptotic variances was written in R, version 2.5.1. Because the model with one parameter to account for variability (the single-σ model) is nested within the model that accounts for variability with multiple parameters (the multiple-σ model in equation (1)), the likelihood ratio test (LRT) can be used to test for significance of the multiple-σ model.

Results

The proposed model in equation (1) was applied to a well-known data set extracted from a developmental toxicity study conducted at the National Toxicology Program (Tyl, et al., 1983) and set $G = 1499$. The study was an experiment whereby fetal implants were injected into 131 CD-1 mice, which were subsequently randomly allocated across *g* = 5 dose levels of the developmentally toxic substance diethylhexyl phthalate (DEHP). The 5 dose levels are 0, 0.025, 0.05, 0.10, and 0.15, in units of % of DEHP in the animal diet. Animals were allocated roughly equally across dose levels. The summarized results of the experiment are in Table 1. The first and last columns of Table 1 are indicative of a threshold dose-response relationship.

The complete vector of parameters for the multiple-σ model applied to this data is given by $(\theta_1,\theta_2,\theta_3,\xi,\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5)$. Because this is a 9-parameter model, the simplex method was applied as described previously. Olsson's algorithm allows up to 20 parameters. The resulting parameter estimates are shown in Table 2 and are compared with the estimates from the original single-σ model. Standard errors (SEs) (also shown in Table 2), were estimated by the method described. Table 2 shows that all related estimates from the two models are relatively comparable, with some noticeable bias in the estimates of the θs. The log-likelihood estimated from (7) is −534.502 with the log-likelihood from the single-σ model being −546.334; the resulting LRT statistic for the test of $H_0 : \sigma_i = \sigma,\ i = 1,...,5$ is 23.664; based on 4 *df*, the p-value is 9.326×10^{-5}, indicating significance of the multiple-σ model.

Figure 1 shows the plot of the linear *B*-spline basis based on equation (3); based on the estimated interior knot value 0.036, the spline basis is constructed to have linearity below and

Table 1: Summary of Results of Study of Fetal Exposure to DEHP

DEHP Dose (% of Animal Diet)	Number of Litters	Total Number of. Fetuses	Number of Affected Fetuses	Proportion of Affected Fetuses
0	30	396	75	0.189
0.025	26	320	37	0.116
0.05	26	319	80	0.251
0.10	24	276	192	0.696
0.15	25	308	302	0.981

above the knot. Figure 2 shows the plot of the logistically transformed estimated dose-response curve based on equation (1) without the random effects. This plot is indicative of two separate linear functions below and above the interior knot.

Figure 3 shows the plots of the two estimated curves from the multiple-σ and single-σ models, respectively. For the multiple-σ model, the interior knot 0.036 is very close to being the threshold value as the below-knot pattern follows closely to a horizontal line. For the single-σ model, it is more of a linear pattern of decreasing slope below the knot. This degree of difference can be important as threshold estimation is crucial aspect of this analysis. Although estimate itself is relatively close, the general dose-response relationship is different and is indicative that the multiple-σ model is more appropriate in this situation.

Conclusion

A linear B-spline threshold dose-response model was modified to include multiple parameters for modeling dose-specific response variation in developmental toxicity study data. The previous model showed only singular response variation across dose groups. Upon application of this new model, it was found that the addition of multiple parameters affected the estimates of non-variation model parameters and led to statistical significance of the new model over the prior one. The spline approach also is more robust than typical approaches that use standard regression functions which restrict the dose-response relationship to one pattern.

Another desirable feature of using splines for fitting is that it includes models which inherently assume a threshold. The approach of using regression splines to account for threshold effects has been used recently in other fields. For example, Molinari, et al. (2001) used spline functions in place of the standard linear functions in a Cox regression analysis of survival data from several clinical trials and Bessaoud, et al. (2005) used spline functions in the logistic regression setting for analysis of clinical data. The spline approach used in this study is similar to these. Both other studies also extended their model to handle several covariates and indicated the practicality of using linear splines to estimate an interior knot as a threshold value. As they were dealing with larger data sets, both groups looked at cases of multiple knots and higher-order spline functions, although neither went past cubic splines and 3

Table 2: Estimates from the Multiple- and Single*-σ Models

Parameter	Multiple-σ Estimates (SE)	Single-σ Estimates (SE)
θ_1	–2.006 (0.352)	–2.022 (0.305)
θ_2	−2.108 (0.277)	−2.530 (0.357)
θ_3	5.519 (0.831)	4.668 (0.490)
ξ	0.036 (0.006)	0.033 (0.007)
σ_1	1.426 (0.266)	1.331 (0.153)*
σ_2	0.009 (0.823)	NA
σ_3	0.783 (0.223)	NA
σ_4	2.554 (0.770)	NA
σ_5	1.947 (0.472)	NA

*Single-σ model has only one variability parameter

Figure 1: The Linear *B*-spline Basis on the Dose Interval [0, 0.15], with Estimated Interior Knot $\hat{\xi} = 0.036$

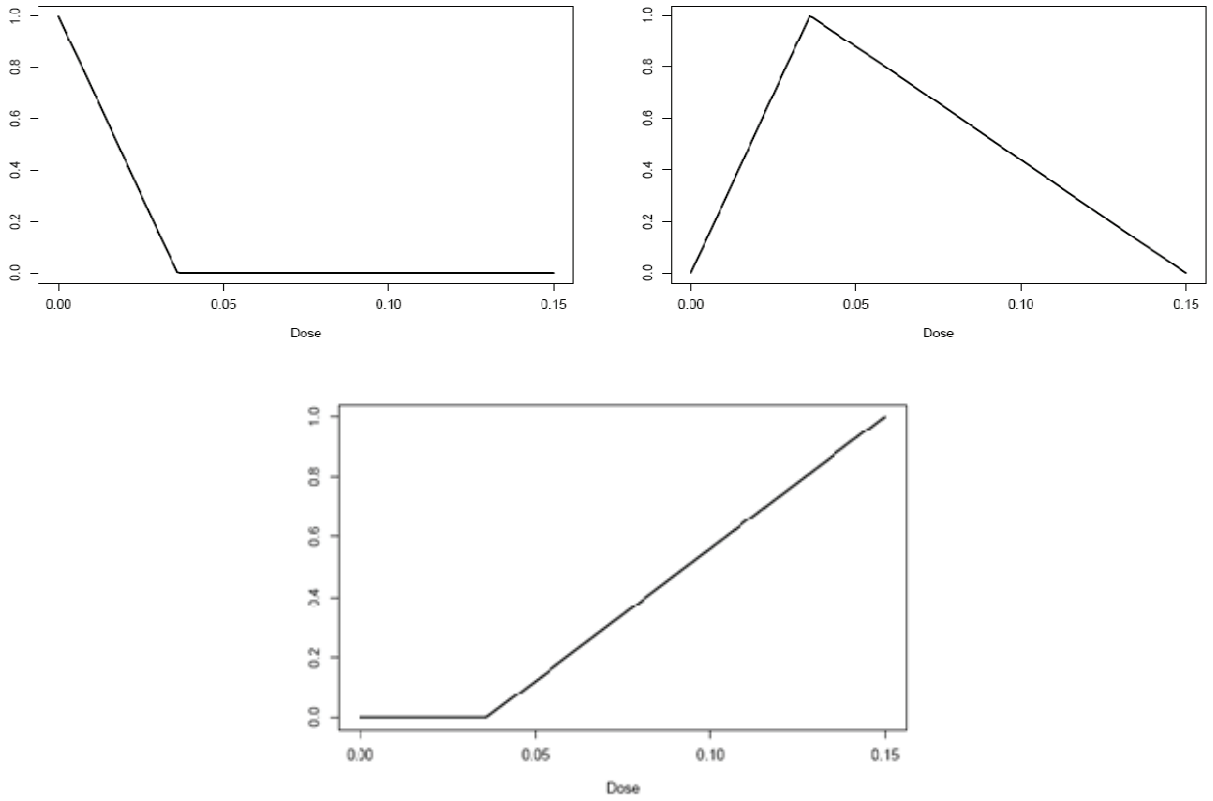


Figure 2: The Fitted Curve (solid line) and 95% Point-Wise Confidence Interval (dashed and dotted lines) for Logit(P)

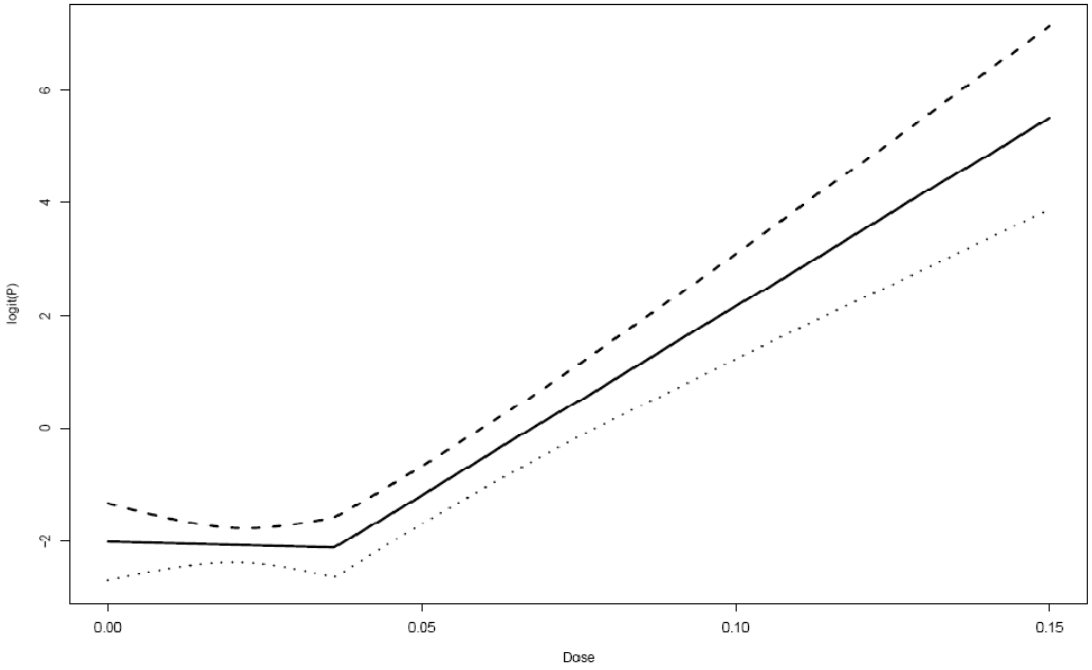
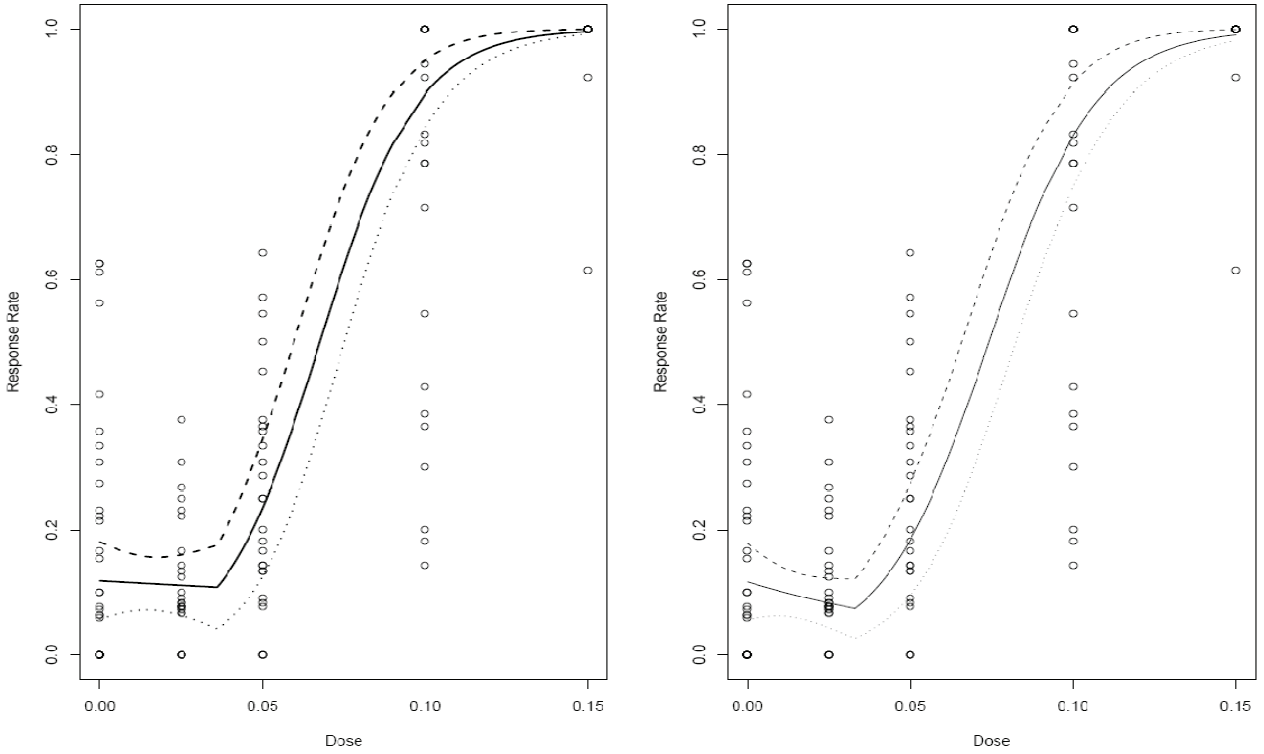


Figure 3: Estimated Dose-Response Curves for the Multiple-σ Model (left) and the Single-σ Model (right)



knots. Bessaoud, et al. (2005) only used up to quadratic splines and both indicated that cubic splines resulted in overfitting and that linear and quadratic seemed appropriate.

For developmental toxicity studies, the existence of a threshold is inherent in current guidelines and is accounted for in some manner during the risk assessment process, albeit indirectly (USEPA, 1991). The pure threshold model is a start in the direction of more adequately modeling these effects, yet the threshold itself has proved to be difficult to ascertain, and any threshold estimated from such a model may be specific to that data set, rather than being a universal value (Cox, 1987). The use of the splines to model behavior that is common to threshold models may help address this issue. Although the interior knot estimated in this study is not specifically the threshold dose level, the model is useful in that it identifies a change point in the direction of the dose-response pattern. It also inherently assumes threshold existence and robustly models several other possible dose-response patterns (Hunt & Li, 2006).

Another advantage of the spline model is the addition of parameters to more adequately model the different degrees of response variation observed to occur across dose groups in a developmental toxicity study. This more accurately specified model improves the estimation of important parameters such as the threshold or, in the case of the spline model, the change point. As illustrated in Kupper, et al. (1986) and in Hunt and Rai (2007), the model assuming multiple parameters to model response variation leads to negligible bias, whereas under conditions of major differences in dose-specific variation the model with single parameter may lead to extremely biased estimates. Thus, the model that has multiple variation parameters is the general model that should be used. However, Hunt and Rai (2007) also showed that in cases of relatively similar variation across dose groups, the single parameter model may suffice.

The potential for future applications in this area include the fitting of higher degree polynomials; for example, the quadratic *B*-spline might be a reasonable extension to the current linear *B*-spline approach. The most immediate advantage is the relative smoothness of the quadratic spline over the linear. However, disadvantages include overfitting. Also, the number of dose levels becomes a factor when adding additional knots into the estimation. The combination of higher order and multiple knots could result in an overly complex model for this type of data. Due to the observed dose-response pattern of the data set under investigation in this article, the linear spline model with one knot appears to provide reasonable fit.

The polynomial regression splines approach is a generally advantageous way to model data from developmental toxicity studies. Rather than requiring a direct estimation of a threshold level, it is able to fit several dose-response curves to the data and implicitly can still indicate the existence of effects such as threshold. It is more robust than previously employed threshold models to such data (Cox, 1987; Schwartz, et al., 1995; Hunt & Rai, 2003, 2007). Additionally, the modification of having dose-specific variation allows for an even more robust model with less biased estimates.

Acknowledgements

This work was supported by Grant Number UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. (C. S. Li). This work was partially supported by Cancer Center Support CA21765 from the National Institutes of Health, USA, and the American Lebanese Syrian Associated Charities (ALSAC) (D. L. Hunt).

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Washington, DC: U.S. Government Printing Office.

Bessaoud, F., Daures, J. P., & Molinari, N. (2005). Free knot splines for logistics models and threshold selection. *Computer Methods and Programs in Biomedicine*, *77*, 1-9.

Collett, D. (1991). *Modelling binary data*. Boca Raton, FL: CRC Press.

Cox, C. (1987). Threshold dose-response models in toxicology. *Biometrics*, *43*, 511-523.

Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology*, *4*, 854-871.

De Boor, C. (2001). *A Practical Guide to Splines* (*Revised Edition*). NY: Springer.

Hunt, D. L., & Li, C. S. (2006). A regression spline model for developmental toxicity data. *Toxicological Sciences*, *92*, 329-334.

Hunt, D. L., & Rai, S. N. (2003). A threshold dose-response model with random effects in teratological experiments. *Communications in Statistics-Theory and Methods*, *32*, 1439-1457.

Hunt, D. L., & Rai, S. N. (2007). Interlitter response variability in a threshold dose-response model. *Submitted to Communications in Statistics-Theory and Methods*.

Kupper, L. L., Portier, C., Hogan, M. D. & Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology. *Biometrics*, *42*, 85-98.

Li, C. S., & Hunt, D. (2004). Regression splines for threshold selection with application to a random-effects logistic dose-response model. *Computational Statistics and Data Analysis*, *46*, 1-9.

Molinari, N., Daures, J. P., & Durand, J. F. (2001) Regression splines for threshold selection in survival data analysis. *Statistics in Medicine*, *20*, 237-247.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308-313.

Olsson, D. M. (1974). A sequential simplex program for solving minimization problems. *Statistical Computer Program*, *6*, 53-57.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*, 425-461.

Ryan, L. (2000). Statistical issues in toxicology. *Journal of the American Statistical Association*, *95*, 304-308.

Schwartz, P., Gennings, C. & Chinchilli, V. (1995). Threshold models for combination data from reproductive and developmental experiments. *Journal of the American Statistical Association*, *90*, 862-870.

Tyl, R. W., Jones-Price, C., Marr, M. C. & Kimmel., C. A. (1983). Teratological evaluation of diethylhexyl phthalate (CAS No. 117-81-7) in CD-1 mice. *Final Study Report for NCTR/NTP Contract NO. 222-80-2031 9(c). NTIS NO. PB85105674*, National Technical Information Service, Springfield, VA.

USEPA. (1991). Guidelines for developmental toxicity risk assessment. *Federal Register*, *56*, 63798-63826.

Appendix

To obtain the observed information matrix $\mathbf{I}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$, the second-order partial derivatives of $\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ in equation (7) must be calculated with respect to $(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ and $\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ as:

$$\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)=\sum_{i=1}^{g}\sum_{j=1}^{m_i}\log(H_{ij}).$$

Here, $H_{ij}=\sum_{l=1}^{q}a_l R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}}$, $R_{il}=\dfrac{\exp(\eta_i+\sigma_i b_l\sqrt{2})}{1+\exp(\eta_i+\sigma_i b_l\sqrt{2})}$, $\eta_i=\sum_{k=1}^{3}\theta_k B_{k,2}(d_i,\xi)=\mathbf{B}_2(d_i,\xi)\boldsymbol{\theta}$, $\bar{R}_{il}=1-R_{il}$, and $\bar{x}_{ij}=n_{ij}-x_{ij}$. The first-order partial derivatives of $\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ with respect to $(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ are:

$$\begin{aligned}\frac{\partial\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial\boldsymbol{\theta}}&=\frac{\partial}{\partial\boldsymbol{\theta}}\sum_{i=1}^{g}\sum_{j=1}^{m_i}\log(H_{ij})\\&=\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\frac{\partial H_{ij}}{\partial\boldsymbol{\theta}}\\&=\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\frac{\partial\eta_i}{\partial\boldsymbol{\theta}}\right\},\end{aligned}$$

$$\begin{aligned}\frac{\partial\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial\sigma_r}&=\frac{\partial}{\partial\sigma_r}\sum_{i=1}^{g}\sum_{j=1}^{m_i}\log(H_{ij})\\&=\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\frac{\partial H_{ij}}{\partial\sigma_r}I_{(r=i)}\\&=\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)b_l\sqrt{2}I_{(r=i)}\right\},\ r=1,\ldots,g,\end{aligned}$$

Appendix (continued)

and

$$\begin{aligned}\frac{\partial \tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial \xi} &= \frac{\partial}{\partial \xi}\sum_{i=1}^{g}\sum_{j=1}^{m_i}\log(H_{ij}) \\ &= \sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\frac{\partial H_{ij}}{\partial \xi} \\ &= \sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\left\{\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+1} - \overline{x}_{ij}R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \xi}\right\},\end{aligned}$$

where

$$\frac{\partial \eta_i}{\partial \boldsymbol{\theta}^T} = \frac{\partial}{\partial \boldsymbol{\theta}^T}\sum\nolimits_{k=1}^{3}\theta_k B_{k,2}(d_i,\xi) = \left[B_{1,2}(d_i,\xi), B_{2,2}(d_i,\xi), B_{3,2}(d_i,\xi)\right] = \mathbf{B}_2(d_i,\xi), \text{ and}$$

$$\frac{\partial \eta_i}{\partial \xi} = \theta_1\frac{d_i - d_1}{(\xi - d_1)^2}I_{(d_i<\xi)} + \theta_2\left(\frac{d_1 - d_i}{(\xi - d_1)^2}I_{(d_i<\xi)} + \frac{d_g - d_i}{(d_g - \xi)^2}I_{(d_i>\xi)}\right) + \theta_3\frac{d_i - d_g}{(d_g - \xi)^2}I_{(d_i>\xi)}.$$

As a result, the components of $\mathbf{I}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)$ are given, respectively, as follows:

$$\begin{aligned}-\frac{\partial^2 \tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i}\left\{\left(\frac{\partial H_{ij}^{-1}}{\partial \boldsymbol{\theta}}\right)\frac{\partial H_{ij}}{\partial \boldsymbol{\theta}^T} + H_{ij}^{-1}\frac{\partial^2 H_{ij}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\} \\ &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\left\{\sum_{l=1}^{q} a_l\left(x_{ij}^2 R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+2} - (x_{ij} + 2x_{ij}\overline{x}_{ij} + \overline{x}_{ij})R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}+1} + \overline{x}_{ij}^2 R_{il}^{x_{ij}+2}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \boldsymbol{\theta}}\frac{\partial \eta_i}{\partial \boldsymbol{\theta}^T}\right\} \\ &\quad + \sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-2}\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+1} - \overline{x}_{ij}R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \boldsymbol{\theta}}\right]\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+1} - \overline{x}_{ij}R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \boldsymbol{\theta}^T}\right],\end{aligned}$$

$$\begin{aligned}-\frac{\partial^2 \tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial \xi\partial \boldsymbol{\theta}^T} &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i}\left\{\left(\frac{\partial H_{ij}^{-1}}{\partial \xi}\right)\frac{\partial H_{ij}}{\partial \boldsymbol{\theta}^T} + H_{ij}^{-1}\frac{\partial^2 H_{ij}}{\partial \xi\partial \boldsymbol{\theta}^T}\right\} \\ &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\frac{\partial^2 H_{ij}}{\partial \xi\partial \boldsymbol{\theta}^T} + \sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-2}\frac{\partial H_{ij}}{\partial \xi}\frac{\partial H_{ij}}{\partial \boldsymbol{\theta}^T} \\ &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\left\{\sum_{l=1}^{q} a_l\left(x_{ij}^2 R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+2} - (x_{ij} + 2x_{ij}\overline{x}_{ij} + \overline{x}_{ij})R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}+1} + \overline{x}_{ij}^2 R_{il}^{x_{ij}+2}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \xi}\frac{\partial \eta_i}{\partial \boldsymbol{\theta}^T}\right\} \\ &\quad - \sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+1} - \overline{x}_{ij}R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial^2 \eta_i}{\partial \xi\partial \boldsymbol{\theta}^T}\right] \\ &\quad + \sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-2}\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+1} - \overline{x}_{ij}R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \xi}\right]\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\overline{R}_{il}^{\overline{x}_{ij}+1} - \overline{x}_{ij}R_{il}^{x_{ij}+1}\overline{R}_{il}^{\overline{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \boldsymbol{\theta}^T}\right],\end{aligned}$$

Appendix (continued)

$$
\begin{aligned}
-\frac{\partial^2 \tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial\sigma_r\partial\sigma_s} &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i}\left\{\left(\frac{\partial H_{ij}^{-1}}{\partial\sigma_r}\right)\frac{\partial H_{ij}}{\partial\sigma_s}+H_{ij}^{-1}\frac{\partial^2 H_{ij}}{\partial\sigma_r\partial\sigma_s}\right\}\\
&= -\sum_{i=1}^{g}\sum_{j=1}^{m_i}\left(\frac{\partial H_{ij}^{-1}}{\partial\sigma_r}\right)\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)b_l\sqrt{2}I_{(s=i)}\right]\\
&\quad -\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a_l\left[\frac{\partial}{\partial\sigma_r}\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\right]b_l\sqrt{2}I_{(s=i)}\right\}\\
&= \sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-2}\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)b_l\sqrt{2}I_{(r=i)}\right]\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)b_l\sqrt{2}I_{(s=i)}\right]\\
&\quad -\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a_l\left[x_{ij}^2R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+2}-\left(x_{ij}+2x_{ij}\bar{x}_{ij}+\bar{x}_{ij}\right)R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}+1}+\bar{x}_{ij}^2R_{il}^{x_{ij}+2}\bar{R}_{il}^{\bar{x}_{ij}}\right]2b_l^2I_{(r=i)}I_{(s=i)}\right\}\\
&= \begin{cases}
-\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a_l\left[x_{rj}^2R_{rl}^{x_{rj}}\bar{R}_{rl}^{\bar{x}_{rj}+2}-\left(x_{rj}+2x_{rj}\bar{x}_{rj}+\bar{x}_{rj}\right)R_{rl}^{x_{rj}+1}\bar{R}_{rl}^{\bar{x}_{rj}+1}+\bar{x}_{rj}^2R_{rl}^{x_{rj}+2}\bar{R}_{rl}^{\bar{x}_{rj}}\right]2b_l^2\right\}\\
\quad +\sum_{j=1}^{m_i}H_{ij}^{-2}\left[\sum_{l=1}^{q}a_l\left(x_{rj}R_{rl}^{x_{rj}}\bar{R}_{rl}^{\bar{x}_{rj}+1}-\bar{x}_{rj}R_{rl}^{x_{rj}+1}\bar{R}_{rl}^{\bar{x}_{rj}}\right)b_l\sqrt{2}\right]^2, & s=r=i\\
0, & s\neq r; s=r\neq i,
\end{cases}
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2 \tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial\xi\partial\sigma_r} &= -\sum_{i=1}^{g}\sum_{j=1}^{m_i}\left\{\left(\frac{\partial H_{ij}^{-1}}{\partial\xi}\right)\frac{\partial H_{ij}}{\partial\sigma_r}+H_{ij}^{-1}\frac{\partial^2 H_{ij}}{\partial\xi\partial\sigma_r}\right\}\\
&= -\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a_l\left(x_{ij}^2R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+2}-(x_{ij}+2x_{ij}\bar{x}_{ij}+\bar{x}_{ij})R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}+1}+\bar{x}_{ij}^2R_{il}^{x_{ij}+2}\bar{R}_{il}^{\bar{x}_{ij}}\right)\sqrt{2}b_lI_{(r=i)}\frac{\partial\eta_i}{\partial\xi}\right\}\\
&\quad +\sum_{i=1}^{g}\sum_{j=1}^{m_i}H_{ij}^{-2}\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\frac{\partial\eta_i}{\partial\xi}\right]\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\sqrt{2}b_lI_{(r=i)}\right]\\
&= \begin{cases}
-\sum_{j=1}^{m_i}H_{ij}^{-1}\left\{\sum_{l=1}^{q}a_l\left(x_{ij}^2R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+2}-(x_{ij}+2x_{ij}\bar{x}_{ij}+\bar{x}_{ij})R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}+1}+\bar{x}_{ij}^2R_{il}^{x_{ij}+2}\bar{R}_{il}^{\bar{x}_{ij}}\right)\sqrt{2}b_l\frac{\partial\eta_i}{\partial\xi}\right\}\\
\quad +\sum_{j=1}^{m_i}H_{ij}^{-2}\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\frac{\partial\eta_i}{\partial\xi}\right]\left[\sum_{l=1}^{q}a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1}-\bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\sqrt{2}b_l\right], & r=i\\
0, & r\neq i,
\end{cases}
\end{aligned}
$$

Appendix (continued)

$$-\frac{\partial^2 \tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\sigma},\xi)}{\partial \xi^2} = -\sum_{i=1}^{g}\sum_{j=1}^{m_i}\left\{\left(\frac{\partial H_{ij}^{-1}}{\partial \xi}\right)\frac{\partial H_{ij}}{\partial \xi} + H_{ij}^{-1}\frac{\partial^2 H_{ij}}{\partial \xi^2}\right\}$$

$$= -\sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\left\{\sum_{l=1}^{q} a_l\left[x_{ij}^2 R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+2} - (x_{ij}+2x_{ij}\bar{x}_{ij}+\bar{x}_{ij})R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}+1} + \bar{x}_{ij}^2 R_{il}^{x_{ij}+2}\bar{R}_{il}^{\bar{x}_{ij}}\right]\left(\frac{\partial \eta_i}{\partial \xi}\right)^2\right\}$$

$$-\sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-1}\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1} - \bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\frac{\partial^2 \eta_i}{\partial \xi^2}\right]$$

$$+\sum_{i=1}^{g}\sum_{j=1}^{m_i} H_{ij}^{-2}\left[\sum_{l=1}^{q} a_l\left(x_{ij}R_{il}^{x_{ij}}\bar{R}_{il}^{\bar{x}_{ij}+1} - \bar{x}_{ij}R_{il}^{x_{ij}+1}\bar{R}_{il}^{\bar{x}_{ij}}\right)\frac{\partial \eta_i}{\partial \xi}\right]^2,$$

where

$$\frac{\partial^2 \eta_i}{\partial \xi^2} = 2\theta_1\frac{d_1-d_i}{(\xi-d_1)^3}I_{(d_i<\xi)} + 2\theta_2\left(\frac{d_i-d_1}{(\xi-d_1)^3}I_{(d_i<\xi)} + \frac{d_g-d_i}{(d_g-\xi)^3}I_{(d_i>\xi)}\right) + 2\theta_3\frac{d_i-d_g}{(d_g-\xi)^3}I_{(d_i>\xi)}$$

and

$$\frac{\partial^2 \eta_i}{\partial \xi \partial \boldsymbol{\theta}^T} = \left[\frac{d_i-d_1}{(\xi-d_1)^2}I_{(d_i<\xi)},\quad \frac{d_1-d_i}{(\xi-d_1)^2}I_{(d_i<\xi)} + \frac{d_g-d_i}{(d_g-\xi)^2}I_{(d_i>\xi)},\quad \frac{d_i-d_g}{(d_g-\xi)^2}I_{(d_i>\xi)}\right].$$

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 583-591

1538 – 9472/09/$95.00

Bayesian Analysis of Evidence from Studies of Warfarin v Aspirin for Symptomatic Intracranial Stenosis

Vicki Hertzberg
Emory University

Barney Stern
University of Maryland

Karen Johnston
University of Virginia

Bayesian analyses of symptomatic intracranial stenosis studies were conducted to compare the benefits of long-term therapy with warfarin to aspirin. The synthesis of evidence of effect from previous non-randomized studies in monitoring a randomized clinical trial was of particular interest. Sequential Bayesian learning analysis was conducted and Bayesian hierarchical random effects models were used to incorporate variability between studies. The posterior point estimates for the risk rate ratio (RRR) were similar between analyses, although the interval estimates resulting from the hierarchical analyses are larger than the corresponding Bayesian learning analyses. This demonstrated the difference between these methods in accounting for between-study variability. This study suggests that Bayesian synthesis can be a useful supplement to futility analysis in the process of monitoring randomized clinical trials.

Key words: Bayesian analysis, Bayesian hierarchical model, Bayesian learning, randomized clinical trial, epidemiology, stroke.

Introduction

A responsibility of the committees charged with monitoring randomized clinical trials is to track new evidence from similar studies. However, there are no specific guidelines for the assembly and analysis of such information. Recently the use of Bayesian methods has become accepted in the randomized clinical trials (RCT) community. (Berry, Berry, McKellar & Pearson, 2003). One area in which Bayesian methods are useful is in the synthesis of evidence. (Spiegelhalter, Abrams & Myles, 2004) Thus such methods could provide a data safety committee with useful insights into relevant external information accumulating during the course of a study.

Vicki Hertzberg is an Associate Professor in the School of Public Health, Department of Biostatistics. Email: vhertzb@sph.emory.edu. Barney Stern is a Professor in the School of Medicine, Department of Neurology and Neurosurgery. Email him at: bstern@som.umaryland.edu. Karen Johnston is a Professor in the Department of Neurology. Email: kj4v@Virginia.EDU.

Although Bayesian methods are growing in acceptance in the RCT community, their use in stroke RCTs is still debated (Berry 2005, Donnan, Davis & Ludbrook, 2005; Howard, Coffey & Cutter, 2005; Krams, Lees & Berry, 2005). This study explores the use of two Bayesian techniques for synthesis of evidence. Specifically sequential Bayesian learning and hierarchical Bayesian models are used (Gelman, Carlin, Stern & Rubin, 2004) to examine results from accumulating studies, then illustrate their application to the Warfarin v Aspirin for Symptomatic Intracranial Disease (WASID) trial.

WASID Background

A long-standing secondary stroke prevention strategy for patients with symptomatic intracranial atherostenosis has been warfarin therapy. Warfarin's use was predicated on evidence published in a case series from the Mayo Clinic in the 1950's (Millikan, Siekert & Shick, 1954). This finding was subsequently supported by similarly positive results in observational studies (Marzewski, et al,. 1982; Moufarrij, Little, Furlan, Williams & Marzewski, 1984; Chimowitz, et al., 1995; Thijs & Albers, 2000; Qureshi, et al., 2003)

In 1998, the National Institute of Neurological Diseases and Stroke (NINDS) funded the Warfarin vs Aspirin for Symptomatic Intracranial Disease (WASID) study, the first double-blinded, placebo-controlled randomized clinical trial (RCT) to test the superiority of warfarin (International Normalized Ratio [INR] 2 – 3) over high-dose aspirin (650 mg twice daily) in this patient population. The protocol called for enrollment of 806 patients with angiographically proven symptomatic intracranial disease to determine a combined endpoint of stroke (ischemic and hemorrhagic) and vascular death. The sample size was chosen to give 80% power to detect a difference between event rates of 33% in the aspirin group compared to 22% in the warfarin group over 3 years after, accounting for a 24% rate of discontinuation of study medications and 1% loss to follow-up, which translates to an alternative hazard ratio (HR) of 1.47.

In July, 2003, after 569 patients were enrolled, NINDS, acting upon advice from the WASID Performance and Safety Monitoring Board (PSMB), stopped WASID because subjects randomized to warfarin were at significant increased risk of major non-endpoint adverse events and the potential for a benefit in primary endpoint events that was sufficient to outweigh these adverse events was very low. Indeed, after study closeout, there was no advantage of warfarin versus aspirin (HR = 1.04; 95% CI = 0.73 to 1.48) (Chimowitz, et al., 2005).

Description of Prior Evidence

Existing literature on warfarin treatment for intracranial stenosis was reviewed (Millikan, et al., 1954; Marzewski, et al., 1982; Moufarrij, et al., 1984; Chimowitz, et al., 1995; Thijs, et al., 2000; Qureshi, et al., 2003). Of the six publications, two studies (Marzewski, et al., 1982; Moufarrij, et al., 1984) insufficiently detailed; focus is placed on the remaining four publications in addition to the article describing the WASID trial results (Chimowitz, et al. 2005). (Relevant features of these studies, along with pertinent effect estimates, are summarized in Table 1.)

Study 1: Millikan, et al. (1954) examined Mayo Clinic patients with either intermittent insufficiency of the basilar system or thrombosis within the basilar arterial system. They found that 10/23 (43%) of patients who did not receive anticoagulant therapy died, compared to 3/21 (14%) of patients receiving anticoagulants. The estimated odds ratio (OR) for death comparing aspirin to warfarin (with associated 95% confidence interval [CI]) is 4.62 (2.18, 9.79).

Study 2: Chimowitz, et al. (1995) assessed cases with symptomatic, angiographically confirmed stenosis (≥ 50%) of a major intracranial artery in a retrospective, non-randomized cohort study. Of the 151 patients included in the study, 88 were treated with warfarin and 63 were treated with aspirin. Treatments and dosages were chosen by local physician. Patients were followed by chart review and telephone or personal / next-of-kin interview until first occurrence of a primary endpoint (major vascular event defined as ischemic stroke, myocardial infarction or sudden death), change in therapy (from aspirin to warfarin or vice versa), or last contact or death due to non-vascular cause. Warfarin patients were followed for a median duration of 14.7 months, experiencing 8.4 major vascular events per 100 patient years of follow-up. Aspirin patients were followed for a median duration of 19.3 months, experiencing 18.1 major vascular events per 100 patient years. The estimate of relative risk (RR) of major vascular events in aspirin patients compared to warfarin patients is 2.2 (95% CI, 1.2, 4.4).

Study 3: Thijs and Albers (2000) interviewed 51 patients identified from chart review. All patients had symptomatic intracranial stenosis and had failed antithrombotic therapy. Of these, 32 patients were followed on warfarin and 19 on aspirin. Cox proportional hazards analysis was conducted to estimate the hazard ratio (HR) for cerebral ischemic events (including TIA) after adjusting for age, presence of anterior circulation disease, Caucasian race, and hyperlipidemia. The estimated aspirin to warfarin HR is 4.9 (95% CI, 1.7, 13.9).

Study 4: Qureshi, et al. (2003) retrospectively assessed 102 patients with symptomatic vertebrobasilar stenosis. Cox proportional hazards analysis gave an estimated

Table 1: Data Used in Study Analyses

Study Number & Author(s)	Year	Endpoint	Warfarin: #events/ #observations	Aspirin: #events/ #observations	Aspirin/Warfarin ratio (95% CI)	Log(ratio) and (sd)	Caveat*
(1) Millikan, et al.	1954	Death	3 / 21 patients	10 / 23 patients	4.62 (2.18, 9.79)	1.53 (0.75)	A
(2) Chimowitz, et al.	1995	Stroke, MI, sudden death	26 / 143 patient-year	14 / 166 patient-year	2.17 (1.16, 4.35)	0.63 (0.33)	B
(3) Thijs and Albers	2000	Cerebral ischemic events	Not given	Not given	4.9 (1.7, 13.9)	0.77 (0.33)	C
(4) Qureshi, et al.	2003	Stroke or death	10 / 619 patient-month	8 / 787 patient-month	0.63 (0.25, 1.59)	-0.46 (0.47)	D
(5) Chimowitz, et al.	2005	Ischemic stroke, brain hemorrhage, vascular death	62 / 504.4 patient-year	63 / 541.7 patient-year	1.04 (0.73, 1.48)	0.06 (0.18)	

Caveats:
A: The treatment received by patients not receiving warfarin is unclear as are the inclusion criteria
B: Retrospective study possibly subject to selection bias
C: HRR is adjusted for age, presence of anterior circulation disease, Caucasian race, hyperlipidemia
D: Unpublished result from data supporting paper; Qureshi & Suri, personal communication, December 22, 2005

HR of 55.6 (95% CI, 9.1, 333), comparing stroke free survival for patients receiving either warfarin or aspirin to patients receiving neither after adjustment for sex, race, hypertension, diabetes mellitus, cigarette smoking, hyperlipidemia, and lesion location. Additional data provided by the authors (Table 2) allowed calculation of the aspirin to warfarin HR as 0.63 (95% CI, 0.25, 1.59).

Study 5: Chimowitz, et al. (2005) was the only RCT comparing warfarin to aspirin in patients with this disease. 569 patients were followed for an average of 1.8 years. The aspirin to warfarin HR is 1.04 (95% CI, 0.74, 1.49).

Table 2: Endpoints (Stroke or Death) in Qureshi, et al. (2003)*

	Warfarin (n=46)	Aspirin (n =40)
Number of Patients	46	40
Stroke or Death	10	8
Person-Months to Endpoint	619	787

*Qureshi & Suri, personal communication, December 22, 2005

Methodology

Bayesian Learning

Because the results of these studies were accumulated over 50 years, a Bayesian learning approach was first used in which the posterior distribution derived from the analysis of the oldest result was used as the prior distribution in order to derive the posterior distribution with the next study. The goal was to estimate the posterior distribution of θ, the unknown mean of the distribution of log(RRR) from its prior and the preceding study results with the posterior distribution derived from study *i-1* serving as the prior distribution for study *i* for $i = 2,\ldots 5$. This is expressed as follows:

Let $Y_i = \log(RRR_i)$.

Assuming that Y_i is a realization from a random distribution depending on θ, the Bayes theorem gives

$$f(\theta|Y_1) \propto f(Y_1|\,\theta) \times f(\theta), \qquad (1)$$

$$f(\theta|Y_2) \propto f(Y_2|\,\theta) \times f(\theta|Y_1). \qquad (2)$$

In general $f(\theta|Y_i) \propto f(Y_i|\,\theta) \times f(\theta|Y_{i-1})$, where $f(\theta)$ is the baseline prior distribution for θ and $i > 1$, assuming that log(RRR) is normally distributed, using the normal distribution for the likelihood and its conjugate, the normal distribution, as the prior for θ.

Hierarchical Random Effects Models

A simultaneous analysis in a hierarchical random effects model was also considered, specifically each has an estimate Y_i of a treatment effect θ_i, such that:

$$Y_i \sim f(y_i \mid \theta_i). \qquad (3)$$

These treatment effects are treated as realizations of random variables from the same population, that is,

$$\theta_i \sim f(\theta_i \mid \theta_\mu), \qquad (4)$$

with θ_μ having its own prior distribution $f(\theta_\mu)$.

Because all of the studies present an estimate of risk which, after transformation, has a normal distribution, a normal distribution was used for functional forms of likelihood and prior distribution functions. For some studies the results may also be viewed as events per person-years of observation per group. In this case Poisson hierarchical models can be used as follows:

> For group j in study i let the number of events, $E_{ij} \sim \text{Poiss}\,(n_{ij} \times \exp(\varphi x_j + \varepsilon_{ij}))$, where x_j is an indicator variable denoting aspirin group membership, $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ and $\varphi \sim N(0, \sigma^2_\varphi)$. Here n_{ij}, the number of person years at risk, is an offset term and φ is the population value for log(RRR).

For each of these analyses a posterior mean was generated with 95% Bayes interval and posterior median with 50% Bayes interval or inter-quartile range.

For each analysis three different baseline priors were used as follows:

1) $\theta \sim N(0, 10)$ (a weakly non-informative prior; warfarin has no effect);
2) $\theta \sim N(0, 0.5)$ (a skeptical prior; warfarin has no effect);
3) $\theta \sim N(0.5, 10)$ (an enthusiastic prior; warfarin reduces risk by 40%).

Additional sensitivity analyses included only studies 2, 4 and 5.

As the Bayesian learning analysis proceeded, graphs of the posterior, likelihood and prior functions were inspected at each step. Thus, the relative influence that the likelihood and prior exerted in determination of the resulting posterior was able to be determined.

Numerical Methods

In addition to the distributions for the parameters of interest, non-informative prior distributions were placed on any nuisance parameters (e.g., σ^2_ε in the hierarchical Poisson model) then integrated over these parameters in the posterior distribution. For estimation, Gibbs sampling (Casella & George, 1992) was used as performed in the WinBUGS software (Spiegelhalter, Thomas, Best, Gilks & Lunn, 2003). Convergence was monitored using the scale reduction factor (SRF) (Gelman, et al., 2004). For each model analyzed, 3 chains were run with 1,000 iterations each (discarding the first 500 in each chain). For analyses which resulted in SRF > 1.1 the number of iterations was increased in each chain by a factor of 10 the

program run again until the SRF ≤ 1.1. Note that such increases were only necessary for analysis of the hierarchical random effects Poisson models.

Results

Bayesian Learning Analyses

The results of the sequential Bayesian learning analysis with log(RRR) as a normal variate using studies 1-4 are shown in Table 3. Note that the Bayesian results that were available at the time that WASID began (studies 1 and 2) were mixed in their support for an effect of warfarin as hypothesized for the WASID clinical trial, i.e., RRR = .33 / .22 = 1.5, versus the null hypothesis RRR = 1.

Specifically, although the 95% Bayes intervals based on the initial informative prior or the initial enthusiastic prior include 1.5 but exclude 1, the interval based on the initial skeptical prior includes both values. With the subsequent addition of study 3's results the evidence favoring warfarin grew stronger. The 95% Bayes intervals stemming from both initial skeptical and initial enthusiastic priors now include 1.5 but exclude 1. Moreover the interval stemming from the initial non-informative prior excludes both 1 and 1.5 to the left. Addition of study 4 has little effect on point and interval estimates. Further point estimates stemming from the non-informative prior tend to be much higher than corresponding point estimates stemming from skeptical and enthusiastic priors at each point. This disparity is due to the difference in variance between the non-informative prior versus skeptical and enthusiastic priors.

A hypothetical future study of warfarin and aspirin would incorporate the results of study 5. With this addition, note that interval estimates stemming from initial non-informative and skeptical priors now include 1. Indeed, the Bayes interval from the initial skeptical prior now excludes 1.5 to the right. The Bayes interval stemming from the enthusiastic prior excludes 1 but covers 1.5 (the alternative hypothesis for WASID) and 2.

Sensitivity analyses including only studies 2, 4 and 5 result in posterior point and interval estimates that are not much different after adding study 4 into the analysis, especially with the skeptical and enthusiastic priors (Table 4). The results after introduction of only study 2 are like the results after inclusion of both studies 1 and 2, suggesting that the optimistic estimates from study 1 do not contribute substantially to the overall conclusion. Additional sensitivity analysis including study 4 produced posterior point and interval estimates that were virtually identical suggesting that study 4 does not have a substantial impact on the analysis.

Table 3: Posterior Point and Interval Estimates for θ = RRR from Bayesian Learning Analysis Using All Studies

Study Added	Interval Type	Non-Informative Prior	Skeptical Prior	Enthusiastic Prior
1	Mean θ (95% Bayes interval)	4.48 (1.14, 17.64)	1.11 (0.75, 1.63)	1.82 (1.23, 2.69)
	Median θ (50% Bayes interval)	4.48 (2.72, 7.39)	1.22 (1.0, 1.35)	1.82 (1.49, 2.23)
2	Mean θ (95% Bayes interval)	2.46 (1.36, 4.44)	1.35 (0.91, 1.99)	1.82 (1.23, 2.69)
	Median θ (50% Bayes interval)	2.46 (2.01, 3.00)	1.35 (1.22, 1.49)	1.82 (1.65, 2.01)
3	Mean θ (95% Bayes interval)	3.00 (1.67, 5.42)	1.65 (1.12, 2.44)	2.01 (1.36, 2.97)
	Median θ (50% Bayes interval)	3.00 (2.46, 3.67)	1.65 (1.49, 1.82)	2.01 (1.82, 2.46)
4	Mean θ (95% Bayes interval)	2.72 (1.65, 4.95)	1.65 (1.11, 2.46)	2.01 (1.35, 3.00)
	Median θ (50% Bayes interval)	2.72 (2.23, 3.32)	1.65 (1.49, 1.82)	2.01 (1.82, 2.23)
5	Mean θ (95% Bayes interval)	1.35 (1.00, 2.01)	1.35 (1.00, 1.43)	1.49 (1.11, 2.01)
	Median θ (50% Bayes interval)	1.49 (1.22, 1.65)	1.35 (1.22, 1.49)	1.49 (1.35, 1.65)

Simultaneous Analysis of RRR Using Hierarchical Random Effects Models with the Normal Distribution

The simultaneous analysis of these studies was examined in the normal model for log(RRR). Posterior point and interval estimates for the analyses of various subsets of studies are shown in Table 5. The results are very similar to the results of the comparable Bayesian learning analysis, although with wider intervals, indicating different consequences of the ways these methods address variability between studies. Specifically the Bayesian learning analysis provides for a posterior variance estimate at each step, but this estimate can drift between steps. In comparison, the simultaneous nature of the hierarchical models requires adjustment over all studies at once.

Simultaneous Analysis of RRR Using Hierarchical Random Effects with the Poisson Distribution

The HR from a Poisson model that compares events per person year between thegroups was examined, with interest in the ratio between the two Poisson parameters, which are an estimate of RRR. Since not all studies were sufficiently detailed in their report of rates, these analyses are limited. Nevertheless the extent of knowledge for 3 of the existing studies and subsets was examined.

The results of these analyses are shown in Table 6. The first analysis, using only study 2, represents a simple Bayesian analysis using a Poisson distribution. Note that the value of 1.5 is included in the 95% Bayes interval estimates, while the null value 1.0 is excluded by the analysis using the non-informative and enthusiastic priors. The other two analyses used a hierarchical random effects Poisson model to adjust for differences between studies. In these analyses using studies 2 and 4 or using studies 2, 4, and 5, the 95% Bayes interval estimates include both 1 and 1.5.

Table 4: Posterior Point and Interval Estimates for θ = RRR from Bayesian Learning Analysis Using Restricted Set of Studies

Study Added	Interval Type	Non-Informative Prior	Skeptical Prior	Enthusiastic Prior
2	Mean θ (95% Bayes interval)	2.23 (1.23, 4.01)	1.35 (0.91, 1.99)	1.82 (1.23, 2.69)
	Median θ (50% Bayes interval)	2.23 (1.82, 2.72)	1.35 (1.22, 1.65)	1.82 (1.65, 2.01)
4	Mean θ (95% Bayes interval)	2.01 (1.22, 3.67)	1.35 (0.90, 2.01)	1.82 (1.22, 2.72)
	Median θ (50% Bayes interval)	2.23 (1.82, 2.46)	1.35 (1.22, 1.49)	1.82 (1.65, 2.01)
5	Mean θ (95% Bayes interval)	1.35 (0.90, 1.82)	1.22 (0.90, 1.65)	1.35 (1.11, 1.82)
	Median θ (50% Bayes interval)	1.35 (1.11, 1.49)	1.22 (1.11, 1.35)	1.35 (1.22, 1.49)

Table 5: Posterior Point and Interval estimates for θ = RRR Using Hierarchical Random Effects Model with Normal Distribution

Studies Included	Interval Type	Non-Informative Prior	Skeptical Prior	Enthusiastic Prior
1, 2, 3, 4	Mean θ (95% Bayes interval)	2.72 (0.27, 14.88)	1.11 (0.67, 1.82)	1.82 (1.22, 3.00)
	Median θ (50% Bayes interval)	3.00 (2.01, 4.06)	1.11 (1.00, 1.35)	1.82 (1.65, 2.23)
1, 2, 3, 4, 5	Mean θ (95% Bayes interval)	2.01 (0.55, 6.69)	1.22 (0.74, 1.82)	1.82 (1.22, 2.72)
	Median θ (50% Bayes interval)	2.01 (1.49, 2.72)	1.22 (1.00, 1.35)	1.82 (1.49, 2.01)
2, 4	Mean θ (95% Bayes interval)	1.49 (0, 4.9×10^5)	1.00 (0.67, 1.65)	1.65 (1.00, 2.72)
	Median θ (50% Bayes interval)	1.82 (0.27, 7.39)	1.00 (0.90, 1.22)	1.65 (1.35, 2.01)
2, 4, 5	Mean θ (95% Bayes interval)	1.22 (0.07, 20.1)	1.11 (0.67, 1.65)	1.49 (1.11, 2.46)
	Median θ (50% Bayes interval)	1.35 (0.90, 2.01)	1.11 (1.00, 1.22)	1.49 (1.35, 1.82)

Conclusion

Reconciliation of the results of WASID with previous reports of a strong effect of warfarin is difficult. Many would advocate that the biases of the prior observational studies should discount those results in favor of the unbiased result of the RCT. Certainly the use of randomization, blinding, standardization of patient management protocols, and central endpoint adjudication ensure bias-free estimate of treatment effect from the RCT. However, RCTs are not without other sources of bias stemming from the selection of participating physicians and clinics as well as the enrollment of consenting patients. Thus, a growing community of investigators (Berry, et al., 2003; Brophy & Lawrence, 1995; Diamond & Kaul, 2004) advocates the use of Bayesian statistical methods to interpret results of clinical trials as well as to synthesize evidence from a set of studies about the effect of treatment(s). Bayesian statistical methods have recently gained notice in the arena of stroke clinical trials (Berry, 2005; Donnan, et al., 2005; Howard, et al., 2005; Krams, et al., 2005).

Although taken as a single trial the WASID results would seem to extinguish the utility of warfarin as a secondary prevention strategy for patients with symptomatic warfarin's demise (Koroshetz, 2005). In this presentation we explore application of Bayesian methods to interpret the WASID results in light of the overall accumulation of evidence regarding the effect of warfarin and consider what insights the Bayesian analyses might have indicated along the way?

At the time of the WASID proposal submission, the accumulated evidence taken from the Bayesian learning perspective fit neatly with the standard of equipoise necessary to justify NIH funding. Specifically, those coming to the debate with no or vague prior beliefs (i.e., the non-informative prior) as well as those favoring warfarin (i.e., the enthusiast) could justify RRR = 1.5 and exclude RRR = 1. On the other hand, those coming to the problem favoring no difference (i.e., the skeptic) could justify both values for RRR. With the hierarchical analyses the alignments of skeptics and enthusiasts remain the same, while those with vague beliefs now align with the skeptics.

In July 2003, when the study was terminated for safety reasons, the results of the Bayesian learning analyses all excluded RRR = 1 from interval estimates, regardless of prior beliefs. When the analysis is restricted to studies meeting perceived quality criteria, the initial state of equipoise described above remained.

Moreover, the hierarchical analyses limited to published results as of July 2003 would be no different than before. However, the inclusion of the rates from study 4, if they had been published at that time, leads to hierarchical model results that lend support for both RRR=1.5 or RRR=1 regardless of prior belief.

The lack of strict correspondence between conclusions from Bayesian learning with those from Bayesian hierarchical random effects models results from differences between methods in incorporating between-study variability. The studies do have differences in design (sample size, endpoint definitions and inclusion criteria) warranting allowances in the modeling process. Although none of studies 1-4 were randomized clinical trials, hierarchical models can be extended to adjust for different classes (such as RCTs versus non-randomized studies) when 2 or more studies of each class are present. Unfortunately only one RCT was available to include.

It is particularly interesting to note the change in conclusions wrought by the unpublished, negative result of Study 4. This finding reinforces the importance of finding all results, even negative ones, in compiling evidence.

The ability to generate interval estimates and use differing priors deepens understanding of the current evidence in light of previous studies. These results point to the utility of Bayesian analyses of prior studies as an additional tool for monitoring clinical trials. The concordance of frequentist and Bayesian efficacy analyses would provide robust confirmation of the appropriateness of a futility analysis when decisions regarding the continuation or stopping of a clinical trial are made.

Acknowledgements

This study was funded by a research grant (1R01 NS33643) from the US Public Health Service National Institute of Neurological Disorders and Stroke (NINDS). We thank Marc Chimowitz, Mike Lynn, Scott Janis, and Bill Powers for their careful review and comments.

References

Berry, D. (2005). Clinical trials: Is the Bayesian approach ready for prime time? Yes! *Stroke*, *36*, 1621-1623.

Table 6: Posterior Point and Interval estimates for φ = RRR Using Hierarchical Random Effects Model with Poisson Distribution

Studies Included	Interval Type	Non-Informative Prior	Skeptical Prior	Enthusiastic Prior
2*	Mean φ (95% Bayes interval)	2.23 (1.22, 4.48)	1.65 (1.0, 3.0)	2.01 (1.22, 3.32)
	Median φ (50% Bayes interval)	2.23 (1.82, 2.72)	1.82 (1.49, 2.01)	2.01 (1.65, 2.46)
2, 4	Mean φ (95% Bayes interval)	0.41 (0, 54200)	1.0 (0.41, 2.72)	1.65 (0.67, 4.48)
	Median φ (50% Bayes interval)	0.41 (0.01, 22.20)	1.0 (0.74. 1.35)	1.65 (1.22, 2.23)
2, 4, 5	Mean φ (95% Bayes interval)	1.11 (0, 24300)	1.0 (0.41, 2.72)	1.65 (0.55, 4.48)
	Median φ (50% Bayes interval)	1.11 (0.03, 54.6)	1.0 (0.74, 1.35)	1.65 (1.11, 2.23)

*Uses simple Poisson model for two groups

Berry, D., Berry, S., McKellar, J., & Pearson, T. (2003). Comparison of the dose-response relationships of 2 lipid-lowering agents: A Bayesian meta-analysis. *American Heart Journal*, *145*, 1036-1045.

Brophy, J., & Lawrence, J. (1995). Placing trials in context using Bayesian analysis: Gusto revisited by reverend Bayes. *JAMA*, *273*, 871-875.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *American Statistician*, *46*, 167-174.

Chimowitz, M., et al. (1995). The warfarin-aspirin symptomatic intracranial disease study. *Neurology*, *45*, 1488-1493.

Chimowitz, M., et al. (2005). Comparison of warfarin and aspirin for symptomatic intracranial arterial stenosis. *New England Journal of Medicine*, *352*, 1305-1316.

Diamond, G., & Kaul, S. (2004). Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. *Journal of the American College of Cardiology*, *43*, 1929-1939.

Donnan, G., Davis, S., & Ludbrook, J. (2005). The Bayesian principle: Can we adapt? *Stroke*, *36*, 1623-1624.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (*2nd Ed.*). NY: Chapman & Hall.

Howard, G., Coffey, C., & Cutter, G. (2005). Is Bayesian analysis ready for use in phase iii randomized clinical trials? Beware the sound of the sirens. *Stroke*, *36*, 1622-1623.

Investigators, T.W.-A.S.I.D.W.T. (2003). Design, progress, and challenges of a double-blind trial of warfarin versus aspirin for symptomatic intracranial arterial stenosis. *Neuroepidemiology*, *22*, 106-117.

Koroshetz, W. (2005). Warfarin, aspirin, and intracranial vascular disease. *New England Journal of Medicine*, *352*, 1368-1370.

Krams, M., Lees, K., & Berry, D. (2005). The past is the future: Innovative designs in acute stroke therapy trials. *Stroke*, *36*, 1341-1347.

Marzewski, D., et al. (1982). Intracranial internal carotid artery stenosis: Longterm prognosis, *Stroke*, *13*, 821-824.

Millikan, C., Siekert, R., & Shick, R. (1954). *Studies in cerrebrovascular disease. Iii. The use of anticoagulant drugs in the treatment of insufficiency or thrombosis within the basilar artery system*. Staff Meetings of the Mayo Clinic, *30*, 116-126.

Moufarrij, N., Little, J., Furlan, A., Williams, G., & Marzewski, D. (1984). Vertebral artery stenosis: Long-term follow-up. *Stroke*, *15*, 260-263.

Qureshi, A., et al. (2003). Stroke-free survival and its determinants in patients with symptomatic vertebrobasilar stenosis: A multicenter study. *Neurosurgery*, *52*, 1033-1040.

Spiegelhalter, D., Abrams, K., & Myles, J. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, England: John Wiley & Sons.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., & Lunn, D. (2003). Bugs: Bayesian *Inference using Gibbs sampling, Technical, MRC Biostatistics Unit*.

Thijs, V., & Albers, G. (2000). Symptomatic intracranial atherosclerosis: outcome of patients who fail antithrombotic therapy. *Neurology*, *55*, 490-498.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 592-593

1538 – 9472/09/$95.00

BRIEF REPORTS

A Maximum Test for the Analysis of Ordered Categorical Data

Markus Neuhäeuser
Koblenz University of Applied Sciences
RheinAhrCampus, Remagen, Germany

Different scoring schemes are possible when performing exact tests using scores on ordered categorical data. The standard scheme is based on integer scores, but non-integer scores were proposed to increase power (Ivanova & Berger, 2001). However, different non-integer scores exist and the question arises as to which of the non-integer schemes should be chosen. To solve this problem, a maximum test is proposed. To be precise, the maximum of the competing statistics is used as the new test statistic, rather than arbitrarily choosing one single test statistic.

Key words: Exact test, maximum test, ordered categorical data, scores.

Introduction

Ordered categorical data occur often in various applications. For example, Gregoire & Driver (1987) pointed out that such ordinal data frequently result from questionnaire surveys in behavioral science investigations. Sheu (2002) noted that ordered categorical variables play an important role in psychological studies because precise measurement is not always possible. Hence, Likert scales are frequently used in psychological research (Rasmussen, 1989). Moreover, ordered categorical data can be found in medical studies (Rabbee, et al., 2003).

When performing exact tests using scores on ordered categorical data, different scoring schemes are possible. In case of three categories the standard scheme is v_1 = (0 0.5 1); because this scheme corresponds to (0 1 2) it is called integer scoring. Ivanova & Berger (2001) proposed non-integer scores: the middle score should be changed to either 0.49 or 0.51 in order to increase the power.

Senn (2007) criticized these non-integer scores because "there is no substantial reason either in terms of likelihood under an alternative hypothesis or on the basis of some other appeal to logic or experience" (p. 297) to replace the standard scheme. Berger (2007) replied that the standard scheme is also arbitrary in case of ordered categorical data and, therefore, the increased power is a rationale for choosing a non-integer scoring scheme. However, the question is which of the two non-integer schemes should be chosen? Berger wrote: "The existence of two viable replacements creates this controversy... If it helps at all, then always shrink to 0, and use only 0.49" (Berger, 2007, p. 299). The latter proposal is arbitrary and may therefore be regarded as unacceptable. However, using the less powerful test with integer scores may also be regarded as unacceptable. Is there an alternative?

In some areas, statistical genetics for example, it is common to apply a maximum test. That is, the maximum of several competing test statistics is used as a new statistic, and the permutation distribution of the maximum is used for inference (Neuhäuser & Hothorn, 2006). Thus, an alternative is using the maximum of the competing statistics as the new test statistic, rather than arbitrarily choosing one single test statistic. Thus, in the case of three categories with the three scoring schemes v_1 = (0 0.5 1), v_2

Markus Neuhäeuser is Professor of Statistics in the Department of Mathematics and Technology. Email him at: neuhaeuser@rheinahrcampus.de.

= (0 0.49 1), and v_3 = (0 0.51 1) the test is performed with the statistic

$$T_{\max} = \max_{i=1,2,3} S(C, v_i)$$

where

$$S(C, v_i) = \frac{\sum_{j=1}^{3} v_{ij} C_{1j}}{n_1} - \frac{\sum_{j=1}^{3} v_{ij} C_{2j}}{n_2}$$

are the individual test statistics with scores v_i = (v_{i1}, v_{i2}, v_{i3}), C_{ij} are the frequencies for ordered category j (j = 1, 2, 3) in group i (i = 1, 2), and n_i is the sample size of group i.

This maximum test has the advantage of a less discrete null distribution and an accompanied increased power as the single tests based on the non-integer scores. The following example was considered by Ivanova & Berger (2001) and discussed by Senn (2007): C_{11} = 7, C_{12} = 3, C_{13} = 2, C_{21} = 18, C_{22} = 4, C_{23} = 14.

In case of this example, the maximum test gives a significant result for the table (C_{11}, C_{12}) = (9, 1), as the scheme v_3 does, in contrast to v_1. For all 76 possible tables with the margins of this example the maximum test's p-value is at least as small as the p-value of the test with the standard scheme. To be precise, the two p-values are identical for 25 tables, but for 51 tables the maximum test's p-value is smaller. Moreover, the maximum test's p-value is always smaller than or equal to the bigger one of the two p-values of the non-integer scoring tests; note that in this example $\max_{i=2,3} S(C, v_i)$ results in an identical test as T_{max}.

Conclusion

The maximum test is a compromise that avoids the arbitrary choice of just one scheme and maintains the advantage of the non-integer scores. Note that the maximum test is not complicated. Because the exact permutation null distribution of T_{max} is used for inference, one does not need to know the correlation between the different $S(C, v_i)$. Thus, when a researcher selects a test based on the trade-off between power and simplicity – as suggested by Ivanova & Berger (2001) – the maximum test is a reasonable choice. In addition, the approach may have some appeal to logic: there is more than one possible test statistic, so combine the competing statistics. Recently, it was shown that a maximum test can be regarded as an adaptive test with the test statistics themselves as selectors (Neuhäuser & Hothorn, 2006). Thus, in a maximum test the data decide and the statistician does not need to arbitrarily choose between different tests.

Note that a multitude of alternative tests applicable to ordered categorical data exist (Liu & Agresti, 2005). However, the discussed score tests as well as the proposed maximum tests have – at a minimum – the advantage that they are easy to apply.

References

Berger, V. W. (2007). [Reply to Senn (2007).] *Biometrics*, 63, 298-299.

Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin*, 101, 159-165.

Ivanova, A., & Berger, V. W. (2001). Drawbacks to integer scoring for ordered categorical data. *Biometrics*, 57, 567-570.

Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14, 1-73.

Neuhäuser, M., & Hothorn, L. A. (2006). Maximum tests are adaptive permutation tests. *Journal of Modern Applied Statistical Methods*, 5, 317-322.

Rabbee, N., Coull, B. A., Mehta, C., Patel, N., & Senchaudhuri, P. (2003). Power and sample size for ordered categorical data. *Statistical Methods in Medical Research*, 12, 73-84.

Rasmussen, J. L. (1989). Analysis of Likert-scale data: a reinterpretation of Gregoire and Driver. *Psychological Bulletin*, 105, 167-170.

Senn, S. (2007). Drawbacks to non-integer scoring for ordered categorical data. *Biometrics*, 63, 296-298.

Sheu, C.-F. (2002). Fitting mixed-effects models for repeated ordinal outcomes with the NLMIXED procedure. *Behavior Research Methods, Instruments, & Computers*, 34, 151-157.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 594-596

1538 – 9472/09/$95.00

An Inductive Approach to Calculate the MLE for the Double Exponential Distribution

W. J. Hurley
Royal Military College of Canada

Norton (1984) presented a calculation of the MLE for the parameter of the double exponential distribution based on the calculus. An inductive approach is presented here.

Key words: MLE, median, double exponential.

Introduction

Norton (1984) derived the MLE using a calculus argument. This article shows how to obtain it using a simple induction argument that depends only on knowing the shape of a function of sums of absolute values. Some introductory mathematical statistics textbooks, such as Hogg and Craig (1970) give the answer to be the median – although correct, this does not tell the whole story as Norton points out; this is emphasized here.

Methodology

It is useful to review the behavior of linear absolute value functions and sums of linear absolute value functions. For example, consider the function

$$g(x) = |1.8 - x|.$$

Its graph is shown in Figure 1. Note that it has a V-shape with a minimum at x = 1.8. Now consider a sum of two linear absolute value terms:

$$h(x) = |1.8 - x| + |3.2 - x|.$$

W. J. Hurley is a Professor in the Department of Business Administration. Email: hurley-w@rmc.ca.

Plots of this function and its components, |1.8 - x| and |3.2 - x|, are shown in Figure 2. Note that $h(x)$ takes a minimum at all points in the interval $1.8 \le x \le 3.2$.

The MLE

The double exponential distribution is given by

$$f(x) = \frac{1}{2} e^{-|x-\theta|}, \quad -\infty < x < \infty.$$

For the sample $\{x_1, x_2, ..., x_n\}$, the log-likelihood function is

$$\ell(\theta) = n \ln(1/2) - \sum_i |x_i - \theta|.$$

Maximizing this function with respect to θ is equivalent to minimizing

$$g_n(\theta) = \sum_i |x_i - \theta|.$$

To obtain the MLE for general n, begin with the case n = 1 where $g_1(\theta) = |x_1 - \theta|$. This function has a minimum at $\theta = x_1$, hence, for n = 1, the MLE is

$$\theta^{MLE} = x_1.$$

Figure 1: Plot of g(x) = | x - 1.8|

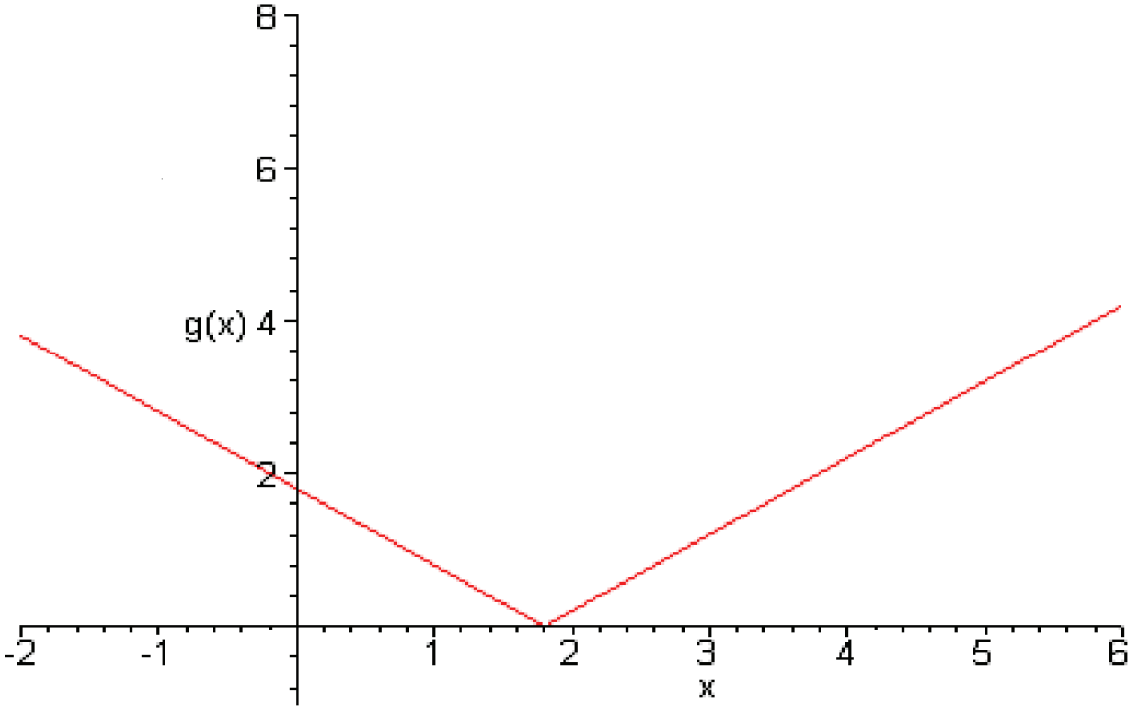
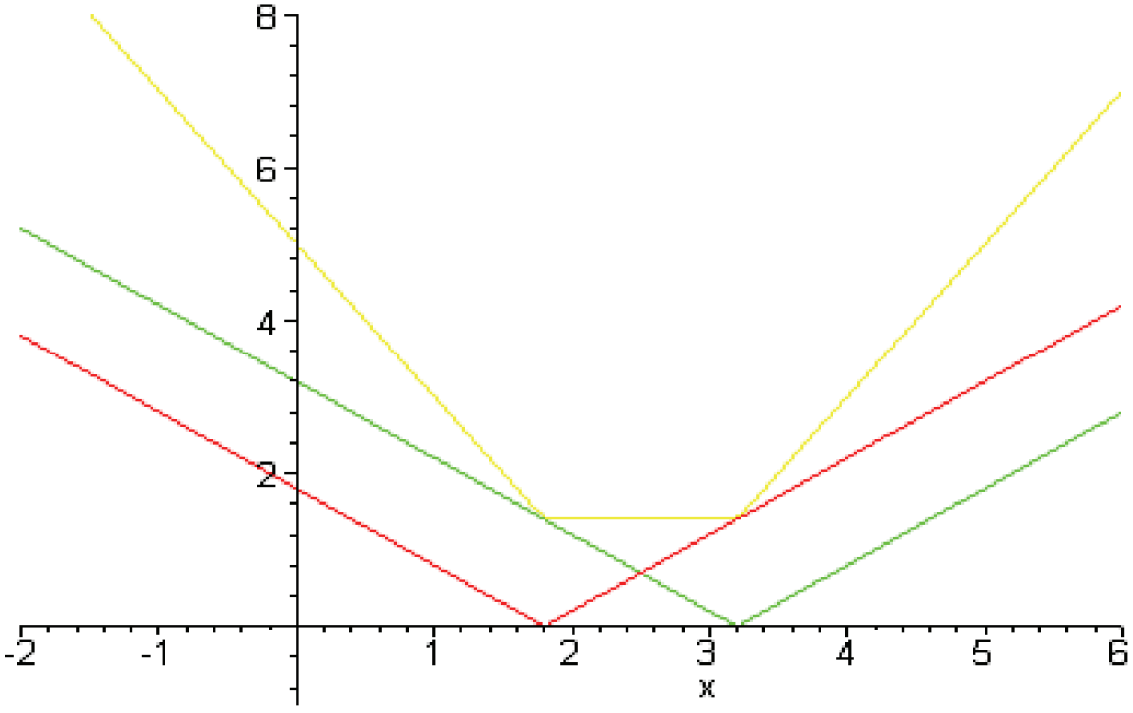


Figure 2: Plot of h(x) = |1.8 - x| + |3.2 - x|



Now, consider the case $n = 2$. For the purposes herein it is useful to order the observations, thus, suppose that the sample is $\{x_{(1)}, x_{(2)}\}$ where $x_{(1)} < x_{(2)}$. The value of θ which minimizes must now be found using

$$g_2(\theta) = |x_{(1)} - \theta| + |x_{(2)} - \theta|.$$

Based on the above, this function takes the form

$$g_2(\theta) = \begin{Bmatrix} -2\theta + x_{(1)} + x_{(2)} & \theta \le x_{(1)} \\ x_{(2)} - x_{(1)} & x_{(1)} \le \theta \le x_{(2)} \\ 2\theta - x_{(1)} - x_{(2)} & \theta \ge x_{(2)} \end{Bmatrix}$$

and has a minimum at any point θ in the interval $x_{(1)} \le \theta \le x_{(2)}$. Hence the MLE for $n = 2$ is

$$\theta^{MLE} = \lambda x_{(1)} + (1-\lambda)x_{(2)}, \quad 0 \le \lambda \le 1.$$

For this case, the median is defined $(x_{(1)} + x_{(2)})/2$ and is a solution, but it is not the only solution.

Next, consider the case $n = 3$ with an ordered sample $x_{(1)} \le x_{(2)} \le x_{(3)}$. Using the same graphical analysis, it can be shown that

$$g_3(\theta) = |x_{(1)} - \theta| + |x_{(2)} - \theta| + |x_{(3)} - \theta|$$

has a unique minimum at $\theta = x_{(2)}$, the median. In the case $n = 4$, the solution is

$$\theta^{MLE} = \lambda x_{(2)} + (1-\lambda)x_{(3)}, \quad 0 \le \lambda \le 1.$$

Thus, the median is a solution, but not the only solution.

Conclusion

Extending the argument for general n is straightforward. It is the median, $x_{((x+1)/2)}$, if n is odd and the generalized median, $\lambda x_{(n/2)} + (1-\lambda)x_{(n/2+1)}$, when n is even.

References

Hogg, R. V., & Craig, A. T. (1970). *Introduction to Mathematical Statistics*, (*3rd Ed.*). New York, NY: MacMillan Publishing Company.

Norton, R. M. (1984). The double exponential distribution: Using calculus to find a maximum likelihood estimator. *The American Statistician*, *38*(*2*), 135-136.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 597-599

1538 – 9472/09/$95.00

New Effect Size Rules of Thumb

Shlomo S. Sawilowsky
Wayne State University

Recommendations to expand Cohen's (1988) rules of thumb for interpreting effect sizes are given to include very small, very large, and huge effect sizes. The reasons for the expansion, and implications for designing Monte Carlo studies, are discussed.

Key words: Effect size, d, Monte Carlo simulation.

Introduction

Some primary considerations for conducting an appropriate Monte Carlo simulation were explicated in Sawilowsky (2003). For convenience, the list is repeated:

- the pseudo-random number generator has certain characteristics (e. g. a long period before repeating values);
- the pseudo-random number generator produces values that pass tests for randomness;
- the number of repetitions of the experiment is sufficiently large to ensure accuracy of results;
- the proper sampling technique is used;
- the algorithm used is valid for what is being modeled; and
- the study simulates the phenomenon in question.

The purpose of this article is to add the following two considerations:

- avoid the use of so-called true random number generators if the randomization process requires replication; and
- ensure study parameters are comprehensive, which necessitates new effect size rules of thumb.

Shlomo Sawilowsky is a WSU Distinguished Faculty Fellow, and Professor of Evaluation and Research. He is the founding editor of *JMASM*. Email: shlomo@wayne.edu.

Regarding the first addition, so-called true random number generators are based on sampling atmospheric or thermal noise, quantum optics, radioactive decay, or other such physical and deterministic phenomena. They aren't seeded, as are pseudo-random number generators, and hence it isn't possible to replicate the sequences they produce. The unscrupulous could make minor substitutions in the sequence to bias the results in such a way that may not be detectable by generic tests for randomness.

Lotteries, military conscriptions, or the like may attempt to overcome this limitation by having the public witness the process via direct observation, which is more compelling than video records that are easily alterable. However, in applications where transparency via replication is essential, such as random sampling in a study commissioned to support allegations in a lawsuit, the use of true random number generators are inappropriate. Thus, if the Monte Carlo study is also a simulation the appropriate number generator, so-called true or pseudo, must be chosen.

Regarding the second addition, Monte Carlo studies conducted on statistical tests' robustness and power properties require choices pertaining to sample sizes, alpha levels, number of tails, choice of competing statistics, inter-correlations of data structures, etc. The study parameters need not, however, be restricted to commonly occurring conditions. In Sawilowsky (1985), the rank transform was studied in the context of a 2×2×2 ANOVA employing sample sizes of 2 to 100 per cell. It is perhaps as unlikely that a classroom or clinic would contain

N=2 study participants as it is that there would be N=100 per cell. Those study parameters were chosen because they represented the minimum and the maximum sample sizes that could be handled given the constraints of the time-share mainframe computing resources available at that time. Prudence dictated sample sizes also be chosen between the two extremes to ensure there were no anomalies in the middle of the robustness rates or power spectrum.

Another important study parameter that must be considered in designing Monte Carlo simulations, which thanks to Cohen (e.g., 1962, 1969, 1977, 1988) has come to be the sin qua non of research design, is the effect size (for an overview, see Sawilowsky, Sawilowsky, & Grissom, in press). Previously, I discussed my conversations with Cohen on developing an encyclopedia of effect sizes:

> I had a series of written and telephone conversations with, and initiated by, Jacob Cohen. He recognized the weaknesses in educated guessing (Cohen, 1988, p. 12) or using his rules of thumb for small, medium, and large effect sizes (p. 532). I suggested cataloging and cross-referencing effect size information for sample size estimation and power analysis as a more deliberate alternative.
>
> Cohen expressed keen interest in this project. His support led to me to delivering a paper at the annual meeting of the AERA on the topic of a possible encyclopedia of effect sizes for education and psychology (Sawilowsky, 1996). The idea was to create something like the "physician's desk reference", but instead of medicines, the publication would be based on effect sizes. (Sawilowsky, 2003, p. 131).

In the context of the two independent sample layout, Cohen (1988) defined small, medium, and large effect sizes as d = .2, .5, and .8, respectively. Cohen (1988) warned about being flexible with these values and them becoming de facto standards for research. (See also Lenth, 2001.) Nevertheless, both warnings are summarily ignored today. That issue cannot be resolved here, but an important lesson that can be addressed is redressing the assumption in designing Monte Carlo studies that the effect size parameters need only conform to the minimum and maximum values of .2 and .8.

For example, when advising a former doctoral student on how to deconstruct the comparative power of the independent t test vs. the Wilcoxon test (Bridge, 2007), it was necessary to model very small effect sizes (e.g., .001, .01). This led to disproving the notion that when the former test fails to reject and the later test rejects it is because the latter is actually detecting a shift in scale instead of a shift in location. It would not have been possible to demonstrate this had the Monte Carlo study began by modeling effect sizes at .2.

Similarly, in the Monte Carlo study in 1985 mentioned above, I modeled what I called a very large effect size equivalent to d = 1.2. This was done because Walberg's (1984) collection of effect sizes pertaining to student learning outcomes included a magnitude of about 1.2 for the use of positive reinforcement as the intervention. Subsequently, in Monte Carlo studies I have conducted, and those conducted by my doctoral students that I supervised, the effect size parameters were extended to 1.2.

As the pursuit of quantifying effect sizes continued even larger effect sizes were obtained by researchers. For example, the use of cues as instructional strategies (d=1.25, Walberg & Lai, 1999), the student variable of prior knowledge (d = 1.43, Marzano, 2000, p. 69), and identifying similarities and differences (d = 1.6, Marzano, 2000, p. 63), exceeded what I defined as very large.

Incredibly, effect sizes on the use of mentoring as an instructional strategy to improve academic achievement have been reported in various studies and research textbooks to be as large as 2.0! The existence of such values, well beyond any rule of thumb heretofore published, has led to researchers presuming the studies yielding such results were flawed.

For example, when DuBois, et al. (2002) were confronted with study findings of huge effect sizes in their meta-analysis of mentoring, they resorted to attributing them as outliers and

deleting them from their study. This was just the first step to ignore the obvious. They then resorted to Winsorizing remaining "large effect sizes [as a] safeguard against these extreme values having undue influence," (p. 167). I have long railed against excommunicating raw data with a large percentage of extreme values as outliers, preferring to re-conceptualize the population as a mixed normal instead of a contaminated normal (assuming the underlying distribution is presumed to be Gaussian; the principle holds regardless of the parent population).

Recently, Hattie (2009) collected 800 meta-analyses that "encompassed 52,637 studies, and provided 146,142 effect sizes" (p. 15) pertaining to academic achievement. Figure 2.2 in Hattie (2009, p. 16) indicated about 75 studies with effect sizes greater than 1. Most fall in the bins of 1.05 to 1.09 and 1.15 to 1.19, but a few also fall in the 2.0+ bin.

Conclusion

Based on current research findings in the applied literature, it seems appropriate to revise the rules of thumb for effect sizes to now define d (.01) = very small, d (.2) = small, d (.5) = medium, d (.8) = large, d (1.2) = very large, and d (2.0) = huge. Hence, the list of conditions of an appropriate Monte Carlo study or simulation (Sawilowsky, 2003) should be expanded to incorporate these new minimum and maximum effect sizes, as well as appropriate values between the two end points.

References

Bridge, T. J. (2007). *Deconstructing the comparative power of the independent samples t test vs the Wilcoxon Mann-Whitney test for shift in location*. Unpublished doctoral dissertation, Detroit, MI: Wayne State University.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology*, *65*, 145-153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*, Rev. Ed. San Diego: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Dubois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of mentoring programs for youth: A meta-analytic review. *American Journal of Community Psychology*, 30 (*2*), 157-197.

Hattie, J. (2009). Visible learning: a synthesis of over 800 meta-analyses relating to achievement. Park Square, OX: Rutledge.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187-193.

Marzano. (2000). A new era of school reform: Going where the research takes us, p. 63. Auroa, CO: Mid-Continent Research for Education and learning.

Sawilowsky, S. (2003a). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*(*1*) 218-225.

Sawilowsky, S. (2003b). A different future for social and behavioral science research. *Journal of Modern Applied Statistical Methods*, *2*(*1*), 128-132.

Sawilowsky, S. S., Sawilowsky, J., & Grissom, R. J. (in press). Effect size. *International Encyclopedia of Statistical Scienc*e. NY: Springer.

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, *41*(*8*), 19-27.

Walberg, H. J., & Lai, J-S. (1999). *Handbook of educational policy*. (G. J. Cizek, Ed.). San Diego, Academic Press, 419-453.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 600-609

1538 – 9472/09/$95.00

Estimation of the Standardized Mean Difference for Repeated Measures Designs

Lindsey J. Wolff Smith S. Natasha Beretvas
The University of Texas at Austin

This simulation study modified the repeated measures mean difference effect size, $d_{RM}^{=}$, for scenarios with unequal pre- and post-test score variances. Relative parameter and *SE* bias were calculated for $d_{RM}^{\neq}$ versus $d_{RM}^{=}$. Results consistently favored $d_{RM}^{\neq}$ over $d_{RM}^{=}$ with worse positive parameter and negative *SE* bias identified for $d_{RM}^{=}$ for increasingly heterogeneous variance conditions.

Key words: meta-analysis, repeated measures, effect size

Introduction

Meta-analysis (Glass, 1976) entails pooling of results from related studies in an effort to synthesize the research results. Studies typically use various experimental designs and thus various effect size measures. In quantitative meta-analysis, a primary goal is to combine effect sizes to produce an overall effect size.

An effect size (ES) index is used to quantify the strength of the relationship between two variables. Each study's finding can be represented as an ES. The use of the ES is important as it allows for the comparison of multiple studies' results. ES indices do, however, differ depending on the type of study performed (e.g., repeated measures, independent groups, etc.). Although multiple effect sizes can be handled using meta-analysis, the effect size of interest in this study is the standardized mean difference for repeated measures designs, δ_{RM}.

Lindsey J. Wolff Smith is a doctoral student in Quantitative Methods in the Department of Educational Psychology. Her interests are in the evaluation of statistical models using simulation research. Email her at: ljwolff@hotmail.com. S. Natasha Beretvas is an Associate Professor and chair of Quantitative Methods in the Department of Educational Psychology. Her interests are in multilevel and meta-analytic modeling techniques. Email her at: tasha.beretvas@mail.utexas.edu.

The formula for the δ_{RM} and its associated variance have been derived by Becker (1988) and Morris and DeShon (2002). The δ_{RM} is necessary for summarizing results from a repeated measures (RM) design in which the same subjects are measured before and after a treatment is administered. Many primary studies employ the RM design. This design allows the researcher to assess change in an outcome that occurs within a subject as a result of what happens between a pre- and post-test. Little research has been done to assess the relative parameter and standard error bias of δ_{RM} estimates.

In the RM design, one group of subjects is measured before and after a treatment is administered. The RM design's ES measure is defined as follows:

$$\delta_{RM} = \frac{\mu_{post} - \mu_{pre}}{\sigma_D} = \frac{\mu_D}{\sigma_D} \qquad (1)$$

where μ_{pre} and μ_{post} are the population means of the pre- and post-test scores, respectively, μ_D is the population mean difference in the pre- and post-test scores, and σ_D is the standard deviation of change scores (Gibbons, Hedeker, & Davis, 1993). The associated sample estimate is calculated as follows:

$$d_{RM} = \frac{\bar{X}_{post} - \bar{X}_{pre}}{s_D}, \qquad (2)$$

where $\overline{X}_{pre}$ and $\overline{X}_{post}$ are the sample means of the pre- and post-test scores, respectively, and s_D is the sample standard deviation of difference scores.

The sampling variance formula for δ_{RM} is:

$$\sigma^2_{\delta_{RM}} = \left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1+n\delta^2_{RM}\right) - \frac{\delta^2_{RM}}{\left[c\left(n-1\right)\right]^2} \quad (3)$$

where n is the number of paired observations in the RM design study (Morris & DeShon, 2002) with a corresponding formula used for sample estimates:

$$s^2_{d_{RM}} = \left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1+nd^2_{RM}\right) - \frac{d^2_{RM}}{\left[c\left(n-1\right)\right]^2}. \quad (4)$$

Equations 3 and 4 also contain the bias correction factor, $c(n-1)$, that is approximated by

$$c\left(n-1\right) = 1 - \frac{3}{4\left(n-1\right)-1} \quad (5)$$

(Hedges, 1982).

Calculation of σ_D is necessary to obtain δ_{RM} (see Equation 1). Morris and DeShon (2002) presented the following relationship between the standard deviation of difference scores, σ_D, and the standard deviation of the original scores, σ:

$$\sigma^{=}_D = \sigma\sqrt{2\left(1-\rho\right)} \quad (6)$$

where ρ is the correlation between the pre- and post-test scores. The corresponding sample estimate is:

$$s^{=}_D = s\sqrt{2(1-r)} \quad (7)$$

with r representing the sample correlation. Both formulas (Equations 6 and 7) are founded on the assumption that the population standard deviations for the pre- and post-test scores are equal (i.e., $\sigma_{pre}=\sigma_{post}=\sigma$). Thus, the notation of including a superscript with = was adopted to distinguish the relevant formula when $\sigma_{pre}=\sigma_{post}$ is assumed from scenarios in which $\sigma_{pre} \neq \sigma_{post}$ is assumed.

If $\sigma_{pre} \neq \sigma_{post}$, another formula for σ_D must be employed that does not assume equal variances, namely:

$$\sigma^{\neq}_D = \sqrt{\sigma^2_{pre} + \sigma^2_{post} - 2\sigma_{pre,post}} \quad (8)$$

where σ^2_{pre} and σ^2_{post} are the population variances of the pre- and post-groups, respectively, and $\sigma_{pre,post}$ is the covariance between the pre- and post-test scores such that:

$$\sigma_{pre,post} = \rho\sigma_{pre}\sigma_{post}. \quad (9)$$

Therefore, the equation for $\sigma^{\neq}_D$ (see Equation 8) becomes:

$$\sigma^{\neq}_D = \sqrt{\sigma^2_{pre} + \sigma^2_{post} - 2\rho\sigma_{pre}\sigma_{post}}. \quad (10)$$

The corresponding sample estimate is then:

$$s^{\neq}_D = \sqrt{s^2_{pre} + s^2_{post} - 2rs_{pre}s_{post}}. \quad (11)$$

Note that when $\sigma_{pre}=\sigma_{post}$, Equations 10 and 11 reduce to the corresponding (population and sample) homogeneous variances formula for σ_D (and s_D) (see Equations 6 and 7, respectively).

This leads to the two primary foci of this study. First, empirical research has not been conducted to assess how well the formulas for δ_{RM} and for $\sigma^2_{\delta_{RM}}$ work in terms of parameter and standard error (*SE*) bias when pre- and post-test scores are and are not homogeneous. Second, applied meta-analysts assume the homogeneity of the pre- and post-test scores and use the $s^{=}_D$ formula (Equation 7) as opposed to $s^{\neq}_D$ (Equation 11) when calculating the estimate of δ_{RM} (Equation 2). Thus, this study also investigated the effect of using the conventional formula for $s^{=}_D$ (Equation 7) when the homogeneity of variance assumption is violated

and the modified formula for s_D (i.e., $s_D^{\neq}$ in Equation 11) should be used.

In the current simulation study, four design factors were manipulated, including: the true value of δ_{RM}, the correlation between pre- and post-test scores, sample size, and values for the pre- and post-test score standard deviations to assess the effect of these factors on parameter and *SE* estimates of δ_{RM}. Results were compared when the pre- and post-test scores were assumed to have equal variances ($\sigma_{pre}^2 = \sigma_{post}^2$), thus $s_D^{=}$ was used to calculate d_{RM} (i.e., providing $d_{RM}^{=}$) with the results based on the assumption that $\sigma_{pre}^2 \neq \sigma_{post}^2$ for which $s_D^{\neq}$ was calculated and used to obtain the associated d_{RM} (i.e., $d_{RM}^{\neq}$).

Methodology

A Monte Carlo simulation study was conducted to assess the relative parameter and *SE* bias of the two estimates of δ_{RM}. The two estimates, $d_{RM}^{=}$ and $d_{RM}^{\neq}$, are distinguished by the formula used to calculate the sample standard deviation of the difference (Equation 7 versus Equation 11). Four design factors were manipulated in this study and are described in detail below. R software version 2.8.1 was used to generate the data and to estimate and summarize all relevant parameters.

δ_{RM}

True values of δ_{RM} were manipulated to assess their effect on parameter and *SE* estimation. These values included: no effect, and small, moderate, and large effects (δ_{RM} = 0, 0.2, 0.5, and 0.8, respectively).

Correlation Between Pre- and Post-Test Scores

The following values of the true correlation, ρ, between pre- and post-test scores were manipulated to evaluate the effect of no, a small, moderate, and large correlation (ρ = 0, 0.2, 0.5, and 0.8, respectively).

Sample Size

Sample size was investigated at three levels including a small, moderate, and moderately large sample size (n = 10, 20, and 60, respectively). Note that the sample sizes used were the same for each of the pre- and post-test groups.

Ratio of the Pre- and Post-Test Scores' Standard Deviations

Five different values of the ratio of the pre- and post-test scores' standard deviations were investigated. The following patterns were evaluated: $\sigma_{pre} = \sigma_{post}$, $\sigma_{pre} < \sigma_{post}$, and $\sigma_{pre} > \sigma_{post}$. For the two unequal standard deviations' conditions, the degree of the difference was also manipulated, with the following four unequal combinations of values for σ_{pre}:σ_{post} investigated: 0.8:1.2, 0.5:1.5, 1.2:0.8, and 1.5:0.5. For the $\sigma_{pre} = \sigma_{post}$ conditions, both pre- and post-test true standard deviations were generated to be one (i.e., $\sigma_{pre} = \sigma_{post} = 1$).

Repeated Measures Effect Size

To manipulate the true value of δ_{RM}, the value of μ_{pre} was set to zero across conditions and the value of μ_{post} was derived to result in the following values for δ_{RM}: 0, 0.2, 0.5, and 0.8. Specifically, μ_{post} is a function of δ_{RM}, μ_{pre}, and σ_D (see Equation 1) and thus can be derived because

$$\mu_{post} = (\delta_{RM})(\sigma_D) + \mu_{pre} \qquad (12)$$

and the values of δ_{RM} and σ_D are determined by the relevant conditions with μ_{pre} always set to zero.

Estimates of δ_{RM}

For each generated dataset, Equation 2 was used to calculate the sample standardized mean difference effect size for RM designs. Two values for s_D ($s_D^{=}$ and $s_D^{\neq}$) were used with the former based on the assumption of equal pre- and post-test score variances (Equation 7) and the latter based on the assumption that $\sigma_{pre}^2 \neq \sigma_{post}^2$ (Equation 11). The resulting estimates were termed $d_{RM}^{=}$ and $d_{RM}^{\neq}$, respectively.

Data Generation

For each set of conditions, a set of random, bivariate normally distributed scores (correlated in the population with the condition's

value for ρ) were generated to provide the pre- and post-test scores for that condition's replication. Two values of d_{RM} ($d^{=}_{RM}$ and $d^{\neq}_{RM}$) were calculated using each dataset as described above. Ten thousand replication datasets were generated for each combination of conditions.

Bias Assessment

Relative parameter and *SE* estimation bias of each d_{RM} ($d^{\neq}_{RM}$ and $d^{=}_{RM}$) was summarized and assessed using Hoogland and Boomsma's (1998) formulas and criteria. More specifically, relative parameter bias was calculated using the following formula:

$$B\left(\hat{\theta}_j\right)=\frac{\left(\overline{\hat{\theta}}_j-\theta_j\right)}{\theta_j} \qquad (13)$$

where θ_j represents the j^{th} parameter's true value and $\overline{\hat{\theta}}_j$ is the mean estimate of parameter j averaged across the 10,000 replications per condition. Hoogland and Boomsma recommended considering a parameter's estimate as substantially biased if its relative parameter bias exceeds 0.05 in magnitude. This cutoff means that estimates that differ from their parameter's true value by more than five percent should be considered substantially biased.

Hoogland and Boomsma's (1998) commonly used formulation of relative standard error bias is as follows:

$$B\left(s_{\hat{\theta}_j}\right)=\frac{\left(\overline{\hat{s}}_{\hat{\theta}_j}-\sigma_{\hat{\theta}_j}\right)}{\sigma_{\hat{\theta}_j}} \qquad (14)$$

where $\overline{\hat{s}}_{\hat{\theta}_j}$ is the mean of the *SE* estimates associated with parameter estimates of θ_j and $\sigma_{\hat{\theta}_j}$ is the empirically true standard error of the distribution of $\hat{\theta}_j$ s calculated by computing the standard deviation of each conditions' 10,000 $\hat{\theta}_j$ s. Hoogland and Boomsma recommended using a cutoff of magnitude 0.10 indicating substantial relative *SE* bias. Note that, for conditions in which the true parameter, δ_{RM}, was zero, simple parameter estimation bias was calculated.

Results

Results are presented in three sections, one for each of the three sample size conditions. Note that relative parameter bias is not calculable if the true parameter value is zero (see Hoogland & Boomsma, 1998), thus, simple bias rather than relative bias is calculated for conditions in which the true δ_{RM} is zero.

Sample Size = 10: Relative Parameter Bias

Substantial positive relative parameter bias was identified for all non-zero values of δ_{RM} and ρ. No substantial bias was found in the $\rho = 0$ conditions. In all cases, the positive bias identified was greater when $d^{=}_{RM}$ was used rather than $d^{\neq}_{RM}$ (see Table 1). No criterion exists to indicate whether simple bias is substantial or not, however, the simple bias values seem small for the $\delta_{RM} = 0$ conditions. When $d^{=}_{RM}$ was used, the more the ratio of $\sigma_{pre} : \sigma_{post}$ values diverged from 1:1, the worse the bias. Similarly, the stronger the ρ, the worse the bias for the $d^{=}_{RM}$ estimate.

The $d^{\neq}_{RM}$ estimator was unaffected by the $\sigma_{pre}:\sigma_{post}$ and ρ values. However, substantial bias was detected for both $d^{\neq}_{RM}$ and $d^{=}_{RM}$ even when $\sigma_{pre}:\sigma_{post}$ was 1:1. Patterns of bias identified for a given $\sigma_{pre}:\sigma_{post}$ ratio closely mimicked patterns identified for the inverse ratio. Thus, across conditions, results found for the 1.5:0.5 ratio matched those for the 0.5:1.5 ratio. Similarly, results for the 0.8:1.2 ratio conditions matched those for the 1.2:0.8 ratio. This result held across all conditions including the three sample sizes and thus will not be mentioned further. Parameter estimation performance of both the $d^{\neq}_{RM}$ and $d^{=}_{RM}$ estimators was unaffected by the true δ_{RM} value (see Table 1). The positive parameter estimation bias of the $d^{\neq}_{RM}$ estimator was pretty consistently close to 10% across the $n = 10$ conditions.

Table 1: Summary of Relative Parameter Estimation Bias by Generating Condition for n = 10 Conditions

	σ_{pre}:σ_{post} Ratio Value									
	1:1		0.8:1.2		0.5:1.5		1.2:0.8		1.5:0.5	
Condition	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$
ρ Value										
0	0.002	0.003	-0.003	-0.003	0.002	0.004	0.002	0.002	0.002	0.004
0.2	***0.090***	***0.106***	***0.097***	***0.130***	***0.096***	***0.199***	***0.093***	***0.125***	***0.087***	***0.190***
0.5	***0.092***	***0.126***	***0.099***	***0.182***	***0.087***	***0.355***	***0.086***	***0.168***	***0.092***	***0.358***
0.8	***0.105***	***0.165***	***0.097***	***0.325***	***0.100***	***0.896***	***0.088***	***0.311***	***0.089***	***0.866***
δ_{RM} Value										
0[a]	0.002	0.003	-0.003	-0.003	0.002	0.004	0.002	0.002	0.002	0.004
0.2	***0.103***	***0.133***	***0.104***	***0.194***	***0.091***	***0.397***	***0.090***	***0.178***	***0.072***	***0.360***
0.5	***0.089***	***0.118***	***0.101***	***0.190***	***0.094***	***0.392***	***0.088***	***0.175***	***0.092***	***0.390***
0.8	***0.092***	***0.121***	***0.093***	***0.181***	***0.093***	***0.394***	***0.088***	***0.176***	***0.096***	***0.399***
Overall[b]	***0.095***	***0.124***	***0.099***	***0.188***	***0.093***	***0.394***	***0.089***	***0.177***	***0.087***	***0.383***

Notes: Substantial relative parameter bias values are highlighted in the table; [a]Mean simple bias is presented for δ_{RM} = 0 conditions; [b] Overall = mean relative parameter bias across all δ_{RM} conditions excluding δ_{RM} = 0 conditions.

Table 2: Summary of Relative Standard Error Estimation Bias by Generating Condition for n = 10 Conditions

	σ_{pre}:σ_{post} Ratio Value									
	1:1		0.8:1.2		0.5:1.5		1.2:0.8		1.5:0.5	
Condition	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$
ρ Value										
0	0.048	0.032	0.046	0.023	0.051	-0.012	0.051	0.027	0.049	-0.011
0.2	0.048	0.062	0.046	0.039	0.051	0.048	0.051	0.053	0.049	0.047
0.5	0.051	0.012	0.041	-0.034	0.046	***-0.147***	0.046	-0.028	0.039	***-0.152***
0.8	0.043	-0.013	0.048	***-0.112***	0.056	***-0.308***	0.047	***-0.110***	0.042	***-0.317***
δ_{RM} Value										
0	0.046	0.016	0.044	-0.031	0.043	***-0.143***	0.042	-0.032	0.038	***-0.145***
0.2	0.041	0.011	0.036	-0.039	0.049	***-0.135***	0.042	-0.030	0.042	***-0.142***
0.5	0.057	0.022	0.047	-0.025	0.049	***-0.134***	0.051	-0.023	0.046	***-0.129***
0.8	0.060	0.020	0.046	-0.027	0.060	***-0.111***	0.061	-0.014	0.052	***-0.120***
Overall[a]	0.051	0.017	0.043	-0.031	0.050	***-0.131***	0.049	-0.025	0.044	***-0.134***

Notes: Substantial relative *SE* bias values are highlighted in the table; [a]Overall = mean relative *SE* bias across all δ_{RM} conditions excluding δ_{RM} = 0 conditions.

Sample Size = 10: Relative *SE* Bias

No relative *SE* bias was found for $d_{RM}^{\neq}$ for the n = 10 conditions (see Table 2). For $d_{RM}^{=}$, however, substantial negative bias was identified in certain conditions. Substantial negative bias (i.e., $\left|B\left(s_{\hat{\theta}_j}\right)\right| > 0.10$, see Equation 14) was found at the most extreme $\sigma_{pre}{:}\sigma_{post}$ values (i.e., when $\sigma_{pre}{:}\sigma_{post}$ = 1.5:0.5 and $\sigma_{pre}{:}\sigma_{post}$ = 0.5:1.5). This bias occurred for conditions in which ρ = 0.5 or larger and the magnitude of the bias seemed to be slightly larger for smaller δ_{RM} (see Table 2). Substantial negative parameter estimation bias was also detected for $d_{RM}^{=}$ for $\sigma_{pre}{:}\sigma_{post}$ = 0.8:1.2 and for $\sigma_{pre}{:}\sigma_{post}$ = 1.2:0.8 for the largest ρ condition (i.e., when ρ = 0.8).

Sample Size = 20: Relative Parameter Bias

No substantial parameter bias was identified when $d_{RM}^{\neq}$ was used to estimate δ_{RM} across the n = 20 conditions (see Table 3).

Substantial positive relative parameter bias was found when $d_{RM}^{=}$ was used to estimate δ_{RM}, however, the degree of parameter bias was lower for the n = 20 conditions (see Table 3) than was observed for the n = 10 conditions (in Table 1).

No substantial relative parameter bias was found in the ρ = 0 conditions for $d_{RM}^{=}$. With the slightly larger sample size, no substantial bias was detected when the $\sigma_{pre}{:}\sigma_{post}$ ratio was 1:1. Otherwise, the pattern of the bias found matched that noted for the n = 10 conditions. The more the value of the $\sigma_{pre}{:}\sigma_{post}$ ratio diverged from 1:1 (and for larger ρ values), the more the degree of substantial parameter bias increased. Values of δ_{RM} did not seem to affect the degree of bias (see Table 3).

Sample Size = 20: Relative *SE* Bias

The relative *SE* bias results for the n = 20 conditions (see Table 4) very closely matched those described for the n = 10 conditions (see Table 2). No substantial relative *SE* bias was found when using $d_{RM}^{\neq}$ to estimate δ_{RM}. For $d_{RM}^{=}$, however, in the most extreme $\sigma_{pre}{:}\sigma_{post}$

Table 3: Summary of Relative Parameter Estimation Bias by Generating Condition for n = 20 Conditions

	$\sigma_{pre}{:}\sigma_{post}$ Ratio Value									
	1:1		0.8:1.2		0.5:1.5		1.2:0.8		1.5:0.5	
Condition	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$
ρ Value										
0	-0.001	-0.001	-0.001	-0.001	0.001	0.002	0.002	0.002	<0.001	<0.001
0.2	0.048	***0.053***	0.038	***0.056***	0.049	***0.120***	0.049	***0.067***	0.038	***0.107***
0.5	0.034	0.046	0.043	***0.097***	0.036	***0.253***	0.043	***0.097***	0.042	***0.258***
0.8	0.038	***0.060***	0.042	***0.219***	0.043	***0.724***	0.039	***0.216***	0.041	***0.722***
δ_{RM} Value										
0[a]	-0.001	-0.001	-0.001	-0.001	0.001	0.002	0.002	0.002	<0.001	<0.001
0.2	0.037	0.047	0.031	***0.094***	0.040	***0.285***	0.040	***0.103***	0.037	***0.283***
0.5	0.043	***0.053***	0.045	***0.110***	0.043	***0.290***	0.048	***0.112***	0.044	***0.288***
0.8	0.045	***0.055***	0.040	***0.103***	0.041	***0.286***	0.042	***0.105***	0.042	***0.288***
Overall[b]	0.041	***0.052***	0.039	***0.102***	0.041	***0.287***	0.043	***0.106***	0.041	***0.287***

Notes: Substantial relative parameter bias values are highlighted in the table; [a]Mean simple bias is presented for δ_{RM} = 0 conditions; [b]Overall = mean relative parameter bias across all δ_{RM} conditions excluding δ_{RM} = 0 conditions.

Table 4: Summary of Relative Standard Error Estimation Bias by Generating Condition for n = 20 Conditions

	σ_{pre}:σ_{post} Ratio Value									
	1:1		0.8:1.2		0.5:1.5		1.2:0.8		1.5:0.5	
Condition	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$
ρ Value										
0	0.020	0.016	0.017	0.007	0.027	-0.006	0.030	0.019	0.010	-0.024
0.2	0.020	0.018	0.017	0.023	0.027	0.010	0.030	0.016	0.010	0.018
0.5	0.017	0.003	0.017	-0.034	0.018	***-0.151***	0.019	-0.031	0.016	***-0.150***
0.8	0.023	0.001	0.011	***-0.118***	0.013	***-0.330***	0.018	***-0.115***	0.018	***-0.329***
δ_{RM} Value										
0	0.018	0.008	0.020	-0.034	0.017	***-0.144***	0.021	-0.034	0.010	***-0.152***
0.2	0.020	0.010	0.010	-0.044	0.016	***-0.144***	0.014	-0.039	0.012	***-0.145***
0.5	0.015	0.003	0.013	-0.041	0.011	***-0.142***	0.021	-0.033	0.019	***-0.135***
0.8	0.025	0.011	0.025	-0.025	0.024	***-0.120***	0.028	-0.026	0.021	***-0.127***
Overall[a]	0.019	0.008	0.017	-0.036	0.017	***-0.138***	0.021	-0.033	0.016	***-0.140***

Notes: Substantial relative *SE* bias values are highlighted in the table; [a]Overall = average relative *SE* estimation bias across δ_{RM} and ρ conditions.

ratio value conditions, substantial negative bias was again found for the stronger ρ conditions (i.e., when ρ = 0.5 and 0.8). The negative relative *SE* bias was slightly worse for smaller δ_{RM} values (see Table 4). Last, substantial negative *SE* bias was also identified for the σ_{pre}:σ_{post} = 0.8:1.2 and σ_{pre}:σ_{post} = 1.2:0.8 conditions in the ρ = 0.8 conditions. Again, slightly worse substantial negative bias was noted for lower true δ_{RM} values.

Sample Size = 60: Relative Parameter Bias

With the larger sample size (n = 60) conditions, the degree of bias decreased further (see Table 5). As with the n = 20 conditions, no substantial bias was detected when $d_{RM}^{\neq}$ was used to estimate δ_{RM}. Substantial positive relative parameter bias was only found in certain conditions when using $d_{RM}^{=}$ to estimate δ_{RM}. Specifically, substantial positive bias was found in the most extreme σ_{pre}:σ_{post} ratio value conditions (i.e., when σ_{pre}:σ_{post} = 1.5:0.5 and σ_{pre}:σ_{post} = 0.5:1.5) and for the ρ = 0.8 conditions when σ_{pre} : σ_{post} = 1.2:0.8 and σ_{pre} : σ_{post} = 0.8:1.2.

The positive bias for ρ = 0.5 paired with the σ_{pre}:σ_{post} = 1.2:0.8 and σ_{pre}:σ_{post} = 0.8:1.2 conditions only just exceeded Hoogland and Boomsma's substantial relative parameter bias criterion. The magnitude of the bias increased for larger ρ values and was unaffected by δ_{RM} values.

Sample Size = 60: Relative *SE* Bias

For the n = 60 conditions, no substantial relative *SE* bias was found with $d_{RM}^{\neq}$ (see Table 6). The same pattern and degree of substantial negative relative *SE* bias as was found for the n = 10 and n = 20 conditions was noted when using $d_{RM}^{=}$ to estimate δ_{RM}. Consistent bias was found for the most extreme σ_{pre}:σ_{post} values when ρ = 0.5 and 0.8 and in the σ_{pre}:σ_{post} = 0.8:1.2 and σ_{pre}:σ_{post} = 1.2:0.8 conditions when ρ = 0.8. The bias was worse within σ_{pre}:σ_{post} values for higher ρ conditions. There seemed to be a very slight effect of δ_{RM} value on the bias with lower δ_{RM} values associated with slightly larger degrees of negative bias (see Table 6).

Table 5: Summary of Relative Parameter Estimation Bias by Generating Condition for n = 60 Conditions

Condition	σ_{pre}:σ_{post} Ratio Value									
	1:1		0.8:1.2		0.5:1.5		1.2:0.8		1.5:0.5	
	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$
ρ Value										
0	<0.001	<0.001	-0.001	-0.001	<0.001	<0.001	-0.001	-0.001	-0.001	-0.001
0.2	0.015	0.017	0.018	0.030	0.007	***0.061***	0.010	0.022	0.010	***0.065***
0.5	0.012	0.016	0.010	***0.053***	0.011	***0.204***	0.013	***0.055***	0.007	***0.200***
0.8	0.015	0.021	0.012	***0.165***	0.010	***0.643***	0.009	***0.162***	0.014	***0.647***
δ_{RM} Value										
0[a]	<0.001	<0.001	-0.001	-0.001	<0.001	<0.001	-0.001	-0.001	-0.001	-0.001
0.2	0.014	0.017	0.010	***0.062***	0.007	***0.229***	0.009	***0.061***	0.013	***0.235***
0.5	0.016	0.019	0.014	***0.066***	0.009	***0.229***	0.012	***0.065***	0.012	***0.232***
0.8	0.013	0.016	0.014	***0.067***	0.013	***0.235***	0.010	***0.062***	0.012	***0.233***
Overall[b]	0.014	0.017	0.013	***0.065***	0.010	***0.231***	0.011	***0.063***	0.012	***0.233***

Notes: Substantial relative parameter bias values are highlighted in the table; [a]Mean simple bias is presented for δ_{RM} = 0 conditions; [b]Overall = mean relative parameter bias across all δ_{RM} conditions except for δ_{RM} = 0 conditions.

Table 6: Summary of Relative Standard Error Estimation Bias by Generating Condition for n = 60 Conditions

Condition	σ_{pre}:σ_{post} Ratio Value									
	1:1		0.8:1.2		0.5:1.5		1.2:0.8		1.5:0.5	
	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$	$d_{RM}^{\neq}$	$d_{RM}^{=}$
ρ Value										
0	0.001	0.001	0.009	0.004	0.004	-0.016	0.015	0.009	-0.003	-0.023
0.2	0.001	0.009	0.009	0.003	0.004	0.004	0.015	0.005	-0.003	0.010
0.5	0.001	-0.003	0.010	-0.030	0.007	***-0.148***	0.001	-0.039	0.009	***-0.146***
0.8	0.014	0.007	0.006	***-0.114***	0.005	***-0.338***	0.013	***-0.109***	0.005	***-0.336***
δ_{RM} Value										
0	0.004	0.001	0.011	-0.036	0.004	***-0.148***	0.008	-0.038	0.008	***-0.142***
0.2	0.006	0.003	0.005	-0.041	0.005	***-0.144***	0.005	-0.042	0.003	***-0.147***
0.5	0.006	0.003	0.010	-0.035	0.012	***-0.132***	0.009	-0.035	0.004	***-0.139***
0.8	0.009	0.005	0.002	-0.040	<0.001	***-0.134***	0.010	-0.033	0.006	***-0.129***
Overall[a]	0.006	0.003	0.007	-0.038	0.005	***-0.140***	0.008	-0.037	0.005	***-0.139***

Notes: Substantial relative *SE* bias values are highlighted in the table; [a]Overall = average relative *SE* estimation bias across δ_{RM} and ρ conditions.

Conclusion

The purpose of this study was to compare estimation of the repeated measures design standardized mean difference effect size, δ_{RM}, using the conventional $d_{RM}^{=}$ estimator with the newly derived $d_{RM}^{\neq}$ modification under a variety of conditions including unequal pre- and post-test score variances. The $d_{RM}^{\neq}$ estimator was designed to correct the standard deviation of the difference scores used in the calculation of δ_{RM} (see Equation 1). The correction recognizes potential differences in the population variances of the pre- and post-test scores. Most statistical tests of differences are based on the assumption that pre- and post-test score variances are equal. However, it is reasonable to assume that this assumption is commonly violated. This study assessed the robustness of the $d_{RM}^{=}$ and $d_{RM}^{\neq}$ estimates of δ_{RM} in scenarios with unequal variances.

Overall, the results convincingly supported use of the suggested modification, $d_{RM}^{\neq}$, as an improved estimator of δ_{RM}. Neither substantial parameter nor *SE* bias was noted for this estimate for sample sizes of 20 or 60 across the spectrum of δ_{RM} and ρ values investigated. In comparison, use of the conventional $d_{RM}^{=}$ estimator, however, cannot be recommended. Substantial positive parameter estimation bias was noted when using the $d_{RM}^{=}$ estimator even in the equal variance conditions (i.e., when $\sigma_{pre}=\sigma_{post}$) for n = 10 and n = 20. Substantial bias was also found across the unequal variance conditions. Negative standard error bias was noted when using the $d_{RM}^{=}$ estimator regardless of sample size. Given the consistency of the degree of *SE* bias across sample sizes of 10, 20, and 60 for the $d_{RM}^{=}$ estimator, it is anticipated that this pattern would be maintained for samples larger than 60.

Substantial parameter bias was identified for the $d_{RM}^{\neq}$ estimator in all of the smallest sample size (n = 10) conditions. (Note that no substantial standard error bias was noted across conditions for the $d_{RM}^{\neq}$ estimator.) The degree of parameter estimation bias in the $d_{RM}^{\neq}$ estimator remained around ten percent across δ_{RM} and ρ values. In other words, the bias was unaffected by the degree of correlation between pre- and post-test scores and by the magnitude of the effect size.

Across conditions, the degree of positive relative parameter bias noted for the $d_{RM}^{=}$ estimator was consistently greater than that noted for the $d_{RM}^{\neq}$ estimator. In addition, the bias detected for the $d_{RM}^{=}$ estimator was affected by the magnitude of ρ. The larger the correlation between pre- and post-test scores, the worse the bias was in the $d_{RM}^{=}$ estimate. The overall degree of positive bias found in the $d_{RM}^{=}$ estimator was greater for smaller sample sizes. But even with samples as large as n = 60, substantial bias was still noted in certain conditions.

The source of the bias noted for the $d_{RM}^{\neq}$ estimator for samples of n = 10 (and not when n was 20 or 60), likely originates in the negative relationship between sample size and degree of bias in the estimation of ρ. Specifically, the conventional estimator, r, (the one used herein) is a biased under-estimate of ρ. Olkin and Pratt (1958) derived an unbiased estimate of ρ, $\hat{\rho}$, that is closely approximated by:

$$\hat{\rho} = r + \frac{r(1-r^2)}{2(n-4)}. \qquad (15)$$

Clearly, the degree of bias exhibited when using r to estimate ρ is represented by $r(1\text{-}r^2)/[2(n\text{-}4)]$ which becomes more substantial with smaller n. Small-sample bias in the estimation of ρ will negatively impact estimation of both $\sigma_D^{\neq}$ (see Equation 8) and $\sigma_D^{=}$ (see Equation 6), ultimately increasing bias in the estimation of δ_{RM} (see Equation 1) for both estimators. Bias in r's estimation of ρ rapidly decreases for larger n which seems to explain the corresponding rapid decrement in the bias of $d_{RM}^{\neq}$'s estimation of δ_{RM}. However, while bias in r's estimation of ρ contributes to the bias noted in $d_{RM}^{=}$'s estimation of δ_{RM}, it cannot fully explain it given $d_{RM}^{=}$'s bias decreases less rapidly than that of $d_{RM}^{\neq}$ for larger n.

Given the consistency in the degree of bias noted for $d_{RM}^{\neq}$ across conditions when n = 10, applied researchers and meta-analysts using $d_{RM}^{\neq}$ as an estimate of δ_{RM} should recognize that, if it is necessary to calculate the repeated measures design standardized mean difference for a sample as small as 10, then it will be over-inflated by about ten percent. Thus, optimally $d_{RM}^{\neq}$ should only be used with sample sizes larger than 10.

Future research should extend this assessment of how well $d_{RM}^{\neq}$ works with smaller sample sizes and should investigate other potential factors that might influence its performance. In addition, future research should extend formulation of the standardized mean difference effect size for repeated measures designs with heterogeneous variances for use with independent groups, repeated measures designs (i.e., for designs with pre- and post-test measures for the treatment and control groups).

A current policy movement encouraging evidence-based practice is leading to an increased use of meta-analysis across the spectrum of medical, educational, and general social science research. Effect sizes summarizing results from studies that have been conducted using repeated measures research designs must also be synthesized to contribute to the evidence base for programs and interventions. While it is commonly assumed that interventions lead to changes in means, not in variances, this is not always the case. This study introduced and validated a correction to the estimate of δ_{RM} that can be used to handle potentially unequal pre- and post-test variances. The new estimator, $d_{RM}^{\neq}$, was found to work better than the conventional one ($d_{RM}^{=}$) across conditions including equal variance conditions. Given the consistently superior performance of $d_{RM}^{\neq}$ over that of the $d_{RM}^{=}$ estimate, applied researchers are encouraged to begin using the $d_{RM}^{=}$ estimator as a less biased estimate of δ_{RM}.

References

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257-278.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics*, *18*, 271-279.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, *5*, 3-8.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490-499.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(*3*), 329-367.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105-125.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201-211.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 610-612

1538 – 9472/09/$95.00

Intermediate r Values for Use in the Fleishman Power Method

Julie M. Smith
Wayne State University

Several intermediate *r* values are calculated at three different correlations for use in the Fleishman Power Method for generating correlated data from normal and non-normal populations.

Key words: Fleishman Power Method, Monte Carlo simulation, correlation.

Introduction

As Headrick and Sawilowsky (1999) observed, "Monte Carlo simulations requiring correlated data from normal and non-normal populations are frequently used to investigate the small sample properties of competing statistics, or the comparison of estimation techniques" (p. 25). Fleishman (1978) introduced the power method for simulating univariate non-normal distributions. This method allows for the systematic control of skew (γ_1) and kurtosis (γ_2) needed in Monte Carlo studies. Fleishman power method models are able to approximate a variety of distributions and require few inputs: a normal random number generator, three constants and an intermediate correlation (Headrick & Sawilowsky, 2000).

A normal random number generator is available as a FORTRAN subroutine, Headrick and Sawilowsky (2000) calculated and provided required constants for various distributions (see Table 1), and the intermediate correlation, *r*, is calculated using the formula

$$r_{xy} = r^2(b^2 + 6bd + 9d^2 + 2a^2r^2 + 6d^2r^4), \quad (1)$$

where *a*, *b* and *d* are constants and r_{xy} is the correlation to which all data will be set. The formula, when solved, results in the graph of a parabola. After graphing, the intermediate *r* value is obtained by determining the point at which the positive horizontal axis intercept is at zero. Establishing the intermediate *r* values may be accomplished via use of a graphing calculator. Values provided in this brief report were obtained using a Texas Instruments (TI) 83-Plus Graphing Calculator by following a five step procedure:

1. Clear all registers and engage the function editor;
2. Enter formula (1) using appropriate constants and desired correlation;
3. Graph the parabola;
4. Use the trace function to position the cursor close to Y = 0 on the positive X-axis;
5. Enlarge the graph using the zoom feature to obtain a precise reading of the positive X value at the point where Y = 0.

Although simple, the process is time-consuming; for this reason several intermediate *r* values have been calculated at three different correlations (See Table 2).

Example

To create correlated data pairs (X, Y) at 0.70 from an exponential distribution (Chi-square, df = 2) with $\gamma_1 = 2$ and $\gamma_2 = 6$, using the constants from Table 1, equation (1) would be as follows:

$$r_{xy} = r^2[(.8263)^2+(6)(.8263)(.02271)+ (9)(.02271)^2+(2)(-.3137)^2r^2+(6)(.02271)r^4]$$

$$r_{xy} = r^2[(.68278)+(.11259)+(.004642)+(.19682)r^2 +(.00309)r^4]$$

Julie M. Smith is a Ph.D. Candidate in the College of Education, Department of Educational Evaluation and Research. Email: ax7955@wayne.edu.

$$r_{xy} = r^2[(.8000)+(.19682)r^2+(.00309)r^4]$$

$$0 = .8000\ r^2+.19682r^4+.00309r^6 - 0.70$$

The positive solution using the stated procedure is r = .859998. This intermediate r value is placed in the following two equations:

$$x_i = rz_1 + \sqrt{1-r^2 z_2} \quad (2)$$

and

$$y_i = rz_1 + \sqrt{1-r^2 z_3} \quad (3)$$

where z_1, z_2 and z_3 are randomly selected standard normal z scores (generated using a random number generator). The data resulting from these equations are not the final correlates, but represent intermediate standard normal variates that will be used to generate the desired correlated data, thus x_i and y_i and the constants appropriate to the distribution are next substituted into the Fleishman equation to produce the final correlates as follows:

$$X = a + bX_i + (-a)X_i^2 + dX_i^2 \quad (4)$$

$$Y = a + bY_i + (-a)Y_i^2 + dY_i^2. \quad (5)$$

The algorithms above produce standardized data centered around $\mu = 0$ and $\sigma = 1$. To realign the values to the χ^2 distribution with df = 2, a simple transformation is performed so that $\mu = 2$ and $\sigma = 2$ as follows:

$$\chi_x^2 = (2)(X) + 2 \quad (6)$$

and for the Y correlate,

$$\chi_x^2 = (2)(Y) + 2 \quad (7)$$

The last step is optional, because computed values are accurate for the distribution. It is only necessary to perform this step if it is desirable to have values commonly found in the tables for the distribution of interest, such as χ^2 (df = 2) in the example.

Table 1: Fleishman Power Constants for Various Distributions*

Distribution	Skew γ_1	Kurtosis γ_2	Constants *a*	*b*	*d*
Chi-square (df=1)	√8	12	−.5207	.6146	.02007
Exponential/Chi-square (df=2)	2	6	−.3137	.8263	.02271
Chi-square (df=3)	1.633	4	−.2595	.8807	.01621
Chi-square (df=4)	√2	3	−.2269	.9089	.01256
Chi-square (df=8)	1	1.5	−.1632	.9531	.0060
Normal	0	0	0	1	0
Cauchy/t (df=1)	0	25	0	.2553	.2038
t (df=3)	0	17	0	.3938	.1713
t (df=7)	0	2	0	.8357	.05206
Laplace/Double Exponential	0	3	0	.7284	.0679

*From Headrick and Sawilowsky (2000), p. 427.

Table 2: Intermediate *r* Values for Various Distributions at Correlations 0.70, 0.80 and 0.90

Distribution	Intermediate *r* Values at Correlation:		
	0.70	0.80	0.90
Chi-square (df=1)	.88909	.92960	.96633
Exponential/Chi-square (df=2)	.85998	.91319	.95973
Chi-square (df=3)	.79989	.85067	.89771
Chi-square (df=4)	.87870	.93855	.99461
Chi-square (df=8)	.84466	.90058	.95271
Normal	.83666	.89443	.94868
Cauchy/t (df=1)	.88121	.92549	.96472
t (df=3)	.86665	.91814	.96118
t (df=7)	.84006	.89697	.95008
Laplace/Double Exponential	.84248	.89877	.95110

References

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, *64*(*1*), 25-35.

Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, *25*(*4*), 417-436.

Sawilowsky, S. S. & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with fortran*. Oak Park, MI: JMASM.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 613-625

1538 – 9472/09/$95.00

Generating and Comparing Aggregate Variables for Use Across Datasets in Multilevel Analysis

James Chowhan Laura Duncan
McMaster University,
Ontario, Canada

This article examines the creation of contextual aggregate variables from one dataset for use with another dataset in multilevel analysis. The process of generating aggregate variables and methods of assessing the validity of the constructed aggregates are presented, together with the difficulties that this approach presents.

Key words: Aggregate variables, contextual variables, multilevel analysis.

Introduction

Contextual effects influence individual outcomes and behaviors. The importance of including community level variables has been gaining ground in the social sciences. Despite their popularity and the presence of theory corroborating the existence of contextual effects, proper measurement and selection of contextual variables continues to challenge researchers. Furthermore, researchers often face the additional difficulty presented by surveys that are not designed to contain contextual information at the geographic area of interest. Even when data is available at the appropriate geographic level, a deficiency of individuals in each area may prohibit the calculation of reliable estimates in multilevel models and thus make it difficult to successfully model contextual effects. A suitable approach to address these difficulties might be to construct aggregate variables in one dataset that has sufficient sample size in the area of interest for use with other datasets.

James Chowhan is a Ph.D. Student in the DeGroote School of Business. Email him at: chowhan@mcmaster.ca. Laura Duncan is a Research Coordinator in Psychiatry and Behavioural Neurosciences in the Offord Centre for Child Studies. Email her at: duncanlj@mcmaster.ca.

Conducting multilevel analyses requires contextual information at the level of interest, for example, family, household, neighborhood, province or country. Datasets are selected by researchers based on their ability to provide answers to research questions and the presence of key variables of interest at the level of interest. In many cases, datasets do not contain the contextual information at the required level for a multilevel analysis. In such cases, researchers could turn to another dataset to construct the desired measure and match this information, using geographical or other identifiers, to their original dataset.

Despite the apparent simplicity of this approach, the issue of checking aggregate variables must be addressed. If possible, the aggregate variables should be checked in some way to assess their validity (i.e., do they measure what they are supposed to measure?). One possible way of checking aggregate variables for validity is presented here, together with problems encountered during the process. These are presented as a means of highlighting some of the hidden complexities of creating aggregate variables that researchers should take into consideration when using this approach.

Methodology

The Survey of Labor and Income Dynamics (SLID) is a longitudinal survey on labor market activity and income. The survey follows individuals with yearly questionnaires

administered for six consecutive years, with a new wave starting every three years since the survey's 1993 initiation. The SLID contains variables that may be used to construct numerous interesting and relevant Economic Region (ER) level variables; thus, researchers could use the following procedures to construct ER level variables of their own choosing. For this example ten aggregate variables on employment and education were constructed; these variables were selected for their potential value to researchers for use in conjunction with other datasets. Table 1 contains the variable names, definitions and the original SLID variables from which they were constructed.

Creation of an Analytic SLID Dataset

First, the variables that were used to create the aggregates (shown in Table 1), were extracted from the SLID together with the appropriate cross-sectional weights and individual and geographical identifiers for each survey year using the SLIDret program. SLID identifies ERs by two separate variables: erres25 and xerres25. The explanation for the presence of two identifiers instead of one is that Statistics Canada amended their ER identification codes in 1999, thus, the SLID contains two sets of ER identification codes. One code refers to the 1991 Census boundaries for all survey years of the SLID (xerres25), and the other refers to the 1991 Census boundaries up to 1999 and to the amended 1999 Census boundaries in subsequent years (erres25). Researchers must decide upon the most appropriate variable to use in any particular research scenario. This will often be determined by the geographical code used in the dataset in which the constructed aggregate variables will be used.

Table 1: Defining Variables of Interest

Variable Name	Definition	SLID Variables
non-employee	Proportion of total labor force self employed	clwrkr1
non-employee_f	Proportion of female labor force self employed	clwrkr1
pct_mgt	Proportion of occupations perceived to be managerial	manag1
pct_mgt_f	Proportion of female occupations perceived to be managerial	manag1
less_hs	Proportion of individuals with less than a high school education	hlev2g18
hs	Proportion of individuals with at least a high school education	hlev2g18
non_univ_ps	Proportion of individuals with a non-university post-secondary certificate	hlev2g18
uni_ps	Proportion of individuals with a university post-secondary certificate	hlev2g18
ps	Proportion of individuals with a post-secondary certificate	hlev2g18
ps(_f)	Proportion of females with a post-secondary certificate	hlev2g18

This article compares the constructed aggregate variables and both the 1996 and 2001 Census profile data. Because the 1996 Census profile data uses the pre-1999 Census boundaries, the xerres25 variable was used to calculate the 1996 SLID ER level estimates. Similarly, because the 2001 Census profile data uses the post-1999 Census boundaries, the erres25 variable was used for the 2001 SLID ER level estimates.

Construction of Aggregate SLID Variables

After creating a SLID dataset, the ER aggregate variables can be constructed. Ten aggregate SLID variables were constructed, seven for the entire population and three for the female population only. The approach was to create a count of individuals in each ER possessing the characteristics of interest and to use this count to construct weighted proportions aggregated at the ER level that could then be exported for use with other datasets.

For each characteristic of interest individuals with that characteristic are dummy coded as 1. This results in dummy variables for individuals aged 15 to 69 who are self employed, individuals aged 15 to 69 whose occupations are perceived as managerial, individuals aged 16 and over who have less than a high school education, individuals aged 16 and over who have at least a high school education, individuals aged 16 and over who have a non-university post-secondary certificate, individuals aged 16 and over who have a university post-secondary certificate, and individuals aged 16 and over who have a post-secondary certificate (university or non-university). There was also a dummy variable for gender so dummy variables could be created for females, for females aged 15 to 69 who are self employed, females aged 15 to 69 whose occupations are perceived as managerial, and females aged 16 and over who have a post-secondary certificate (university or non-university).

Aggregating SLID to the ER Level

After creating SLID dummy variables; the final step was to aggregate these variables. In all cases these aggregates will be proportions for each ER created by aggregating up to the ER level. Because the SLID produces an annual cross-section of individuals it is also necessary to aggregate to the ER level by survey year in order to obtain an accurate estimate of area level characteristics for each year. Taking the mean of a dummy variable is one way to calculate the proportion of individuals with a certain characteristic. Hence, proportions for each ER are calculated by collapsing the dummy variables to their mean for each ER level and for each survey year. These proportions are weighted using the cross-sectional weight. The resulting aggregate variables represent proportions of individuals in ERs with the characteristics of interest outlined.

Once created, aggregates are ready for use; however, it is highly recommend that a check be carried out to assess their validity as aggregate measures. This is accomplished in the following example by comparing the provincial and national population totals followed by the basic gender and age characteristics of the samples. The logic being that, if the population totals are similar and sample characteristics are similar across these demographics, there is some reason to assume that they will be similar in other ways. It is not guaranteed that this is actually the case, however.

As an additional check, similar education and employment aggregates constructed using the Census profile data from 1996 and 2001 were compared as well. (This will not be an option readily available to researchers if one of the main reasons for going to another dataset is that the variables of interest are not available in the Census profile data.) These comparisons are recommended because they will alert a researcher to oddities about the variables or dataset used and to inconsistencies that may require investigation.

To assess the validity of the aggregate SLID variables constructed, a comparison was made to the 1996 and 2001 Census. The Census is, by design, the most accurate and representative approximation of true population parameters. In order for the SLID aggregates to be useful they should reflect true population parameters. It may be argued that using the Census to verify how closely the SLID data and constructed aggregates reflect the true population is the most suitable method of comparison available. As the SLID weighting is

calibrated on Census population totals, it is expected that estimates will match well. The following is a step-by-step guide to comparing aggregate variables.

Choose a Method of Comparison

Two methods of comparison were used in this example. The first involved simply calculating and comparing provincial and national population totals for both SLID datasets for 1996 and 2001. If no similarity existed at this level it would not be sensible to continue with the comparison and the validity of the aggregate variables would be questionable. The second method of comparison used confidence intervals as a means of statistically assessing how close the estimates match. This requires similarly defined variables to be created using Census profile data so that aggregates are created from the SLID and the Census profile data at the ER level. Confidence intervals (assuming a Normal distribution) can be created around the SLID estimates and observations made as to whether the population estimates from the Census fall within these confidence intervals for each ER. The confidence level chosen for this example is 95% but researchers can choose any level they think is suitable. A high number of matches show the SLID estimates are a good match to true population parameters.

Choose and Generate Demographic Variables and Confidence Intervals

For the provincial and national population totals, weighted sums were calculated in STATA broken out by province. At the ER level, two characteristics were chosen for comparison: gender and age. Twenty-one age and gender breakouts by ER were calculated using the SLID data for 1996 and 2001. In addition to the proportion of females, age breakouts for the whole population and for females only are generated using different age intervals. Using STATA, 95% confidence intervals were created for each SLID estimate.

Recreate Aggregate Variables Using Census Profile Data

Because the Census profile data does not contain ER identification codes it is first necessary to merge the Census data with the Postal Code Conversion File (PCCF), matching the data by Enumeration Area (EA) for 1996 and Dissemination Area (DA) for 2001. Enumeration areas (EA) in the 1996 Census and Dissemination areas (DA) in the 2001 Census are smaller geographical areas making up various larger Statistics Canada geographical areas, including ERs. The 1999 change in Census boundaries lead to a name and definition change from EA to DA (for more information on using the PCCF and the change from EA to DA see Gonthier, et al., 2006).

For 1996 the EA code is an eight-digit code constructed from provincial, federal and EA identifiers. The provincial code composes the first two digits; the federal code the following three and the EA code the final three. To construct the eight-digit EA code from its composite parts the provincial code is multiplied by 1,000,000 and the federal code is multiplied by 1,000, and these numbers are added to the EA code. The Census data is then merged with the 1996 PCCF file using this eight-digit EA identifier.

For 2001 the DA code is an eight-digit code constructed from provincial, census division and DA identifiers. The provincial code composes the first two digits, the census division code the following two and the DA code the final four. The eight-digit DA code for 2001 is created the same way as the EA code for 1996. Merging results in each record being assigned an ER identification code. Once again, to ensure the production of accurate estimates, data is aggregated to the ER level by first creating a sum of all individuals within ERs with the characteristics of interest. This ensures accurately weighted estimates reflecting the numbers of individuals in ERs. After these sums are created for each ER proportions are then calculated that correspond to the ten aggregate variables created in the SLID. Table 2 shows the variable names and the 1996 and 2001 Census variables from which they were constructed.

Compare SLID and Census Profile Data Estimates

With the SLID education and employment aggregates, population totals, demographic variables, confidence intervals for these estimates for 1996 and 2001 and similar

Table 2: Concordance between 1996 and 2001 Census Variables

Variable Name	Definition	1996 Census Variable Range Used	2001 Census Variable Range Used
non_employee	Proportion of total labor force self employed	v1211-v1222	v949-v960
non_employee_f	Proportion of female labor force self employed	v1235-v1246	v973-v984
pct_mgt	Proportion of occupations perceived to be managerial	v1031-v1090	v985-v1044
pct_mgt_f	Proportion of female occupations perceived to be managerial	v1151-v1210	v1105-v1164
less_hs	Proportion of individuals with less than a high school education	v1338-v1351	v1382-v1395
hs	Proportion of individuals with at least a high school education	v1338-v1351	v1382-v1395
non_univ_ps	Proportion of individuals with a non-university post-secondary certificate	v1338-v1351	v1382-v1395
uni_ps	Proportion of individuals with a university post-secondary certificate	v1338-v1351	v1382-v1395
ps	Proportion of individuals with a post-secondary certificate	v1352-v1375	v1358-v1381
ps_f	Proportion of females with a post-secondary certificate	v1352-v1375	v1358-v1381

variables recreated using Census profile data from 1996 and 2001, the comparison was carried out. First, weighted provincial and national population totals were compared by year and by province; results are shown in Tables 3a and 3b.

It is important to note that some variation in the totals is to be expected due to rounding error in the Census. In both tables it was expected that column 1 and 2 add up to column 3. In 1996, there was a difference of 685 and in 2001 there is a difference of 235. These differences are likely due to rounding error. It would also be expected that column 4 and column 6 would be similar and that column 5 would be less than both of these. In 1996 the total population is SLID is 271,963 more than the Census total population and in 2001, the total population in the SLID is 1,828,145 below the Census total population: no obvious reason exists to explain this. Even with the minor disparity, population totals in the SLID are close enough to the Census to conclude that the data matches reasonably well.

Second, basic demographics were compared by year and by ER in order to determine the number of Census profile estimates that would fall within the 95% confidence intervals generated around the SLID estimates. Each Census profile estimate falling within the confidence interval was called a match. Table 4 shows the percentage of matches across 66 ERs in 1996 and 73 ERs in 2001.

The proportion of females in the population variable matched perfectly and the age breakouts had a high, but not perfect, percentage of matches. The only variables with suspiciously low numbers of matches were the percentage of individuals aged 15 to 19 and the percentage of females aged 15 to 19. The age breakouts for individuals and females aged 15 to 25 and 20 to 24 showed much better matching. This suggests that the discrepancy is occurring at

the lower end of the age spectrum in the 15 to 19 age range.

Based on observations of similarities in the population totals, gender and age characteristics across the SLID and Census profile data samples, it may be suggested that the SLID and the Census profile data will also be similar across other characteristics, in this case education and employment. To test this, the constructed aggregates were checked for validity in a similar manner. Again, 95% confidence intervals (assuming a Normal distribution) were created around the SLID estimates and observations were made as to whether the population estimates from the Census fell within these confidence intervals for each ER. Table 5 shows the percentage of matches across 66 ERs in 1996 and 73 in 2001.

Given the excellent age and gender match of the data, the low number of matches for the constructed aggregate variables is surprising. Without a clear explanation as to why the variables do not match, the constructed aggregates cannot be trusted as representative and should not be used. However, if explanations can be found for the low matching then the aggregates may be of some use. An investigation of the data and variable definitions was carried out to identify possible causes for the low number of matches.

Investigation of the data and examination of the documentation highlighted several limitations with the variables chosen for use in both the Census profile data and the SLID. These limitations are very likely the cause of the low number of matches across the aggregate variables. First, the internal consistency of the constructed estimates was investigated. In particular, confirmation was required that the total populations being used on the SLID and in the Census Profile data as the denominator in the proportions calculations were in fact the sum of their composite parts. In both the SLID and the Census Profile data, age and education populations were verified. A check was made of the proportions of individuals aged under 25, 25 to 49, 50 to 74 and 75; these proportions should total 1 as this range of ages encompasses all possible ages in the population. The same check was carried out for the female proportions and for the proportions of individuals aged under 25, 25 to 49, 50 to 64 and

Table 3a: Provincial and National Totals for SLID and Census Profile Data, 1996

1996	Census					SLID
Province	1. Male Subtotal	2. Female Subtotal	3. Total Population	4. Total Population 15+	5. Total Labor Force 15+	6. Total Population 15+
10	271,740	278,575	550,420	435,985	245,165	423,747
11	65,990	68,450	134,440	103,580	70,695	100,100
12	441,490	466,175	907,635	718,015	438,010	669,414
13	362,490	374,665	737,255	583,550	363,055	556,031
24	3,318,800	3,462,665	6,781,570	5,382,325	3,357,080	5,394,101
35	4,794,345	5,011,300	9,805,685	7,669,850	5,084,190	7,848,826
46	492,640	509,980	1,002,730	769,900	511,145	782,124
47	450,690	461,550	912,085	689,015	463,360	687,939
48	1,225,800	1,227,510	2,453,330	1,864,640	1,348,880	1,895,376
59	1,706,985	1,751,340	3,458,715	2,743,105	1,819,185	2,874,269
Total	13,130,970	13,612,210	26,743,865	20,959,965	13,700,765	21,231,928

Table 3b: Provincial and National Totals for SLID and Census Profile Data, 2001

2001	Census					SLID
Province	1. Male Subtotal	2. Female Subtotal	3. Total Population	4. Total Population 15+	5. Total Labor Force 15+	6. Total Population 15+
10	249,805	260,815	510,545	422,170	240,600	404,336
11	65,450	69,145	134,530	107,940	73,570	98,323
12	437,335	466,330	903,505	739,060	450,075	681,910
13	355,380	371,485	726,990	597,500	370,920	548,849
24	3,521,985	3,689,680	7,212,255	5,923,010	3,734,615	5,270,975
35	5,458,005	5,701,920	11,159,880	8,972,500	5,950,800	8,426,920
46	547,455	567,110	1,114,400	881,395	582,590	796,246
47	478,785	494,380	973,075	766,390	509,670	691,965
48	1,470,895	1,473,690	2,944,620	2,334,465	1,678,965	2,193,306
59	1,908,975	1,978,245	3,887,305	3,183,715	2,046,190	2,994,323
Total	14,494,070	15,072,800	29,567,105	23,928,145	15,637,995	22,100,000

Table 4: Comparison of SLID and Census Profile Aggregate Estimates for Gender and Age Variables

Variable Name	Variable Definition	1996	2001
		% of Matches	% of Matches
female	% population that is female	100	97
pct_15to25	% population aged 15 to 25	71	70
pct_25to49	% population aged 25 to 49	79	78
pct_50to74	% population aged 50 to 74	70	84
pct_75over	% population aged 75 & over	62	78
pct_50to64	% population aged 50 to 64	68	79
pct_65over	% population aged 65 & over	70	74
pct_15to19	% population aged 15 to 19	44	52
pct_20to24	% population aged 20 to 24	80	86
pct_40to44	% population aged 40 to 44	90	88
pct_75to79	% population aged 75 to 79	74	89
pct_15to25_f	% female population aged 15 to 25	68	74
pct_25to49_f	% female population aged 25 to 49	85	88
pct_50to74_f	% female population aged 50 to 74	76	86
pct_75over_f	% female population aged 75 & over	73	88
pct_50to64_f	% female population aged 50 to 64	79	86
pct_65over_f	% female population aged 65 & over	77	82
pct_15to19_f	% female population aged 15 to 19	70	66
pct_20to24_f	% female population aged 20 to 24	80	92
pct_40to44_f	% female population aged 40 to 44	85	92
pct_75to79_f	% female population aged 75 to 79	83	92

Table 5: Comparison of SLID and Census Profile Aggregate Estimates for Employment and Education Variables

Variable Name	Variable Description	% of Matches	
		1996	2001
non_employee	Proportion of total labor force self employed	44	55
non_employee_f	Proportion of female labor force self employed	61	67
pct_mgt	Proportion of occupations perceived to be managerial	20	8
pct_mgt_f	Proportion of female occupations perceived to be managerial	30	22
less_hs	Proportion of individuals with less than a high school education	33	51
hs	Proportion of individuals with at least a high school education	2	0
non_univ_ps	Proportion of individuals with a non-university post-secondary certificate	42	79
univ_ps	Proportion of individuals with a university post-secondary certificate	35	60
ps	Proportion of individuals with a post-secondary certificate	74	59
ps_f	Proportion of females with a post-secondary certificate	88	85

65 and over. It was found that the 15 to 19 age category produced low numbers of matches. Verification was made that the difference between the proportion of individuals aged under 25 and the proportion of individuals aged 15 to 19 added to the proportion of individuals aged 20 to 24 equals 0. The same verification was made for the female proportions. The results were either extremely close or exactly 0 or 1 (See Appendix 1 for details).

Additional checks were carried out for the education variables in the Census Profile data for both 1996 and 2001. The difference between the proportion of individuals with postsecondary certificates and the proportion of individuals with university certificates added to the proportion of individuals with non-university certificates was checked with the expectation that, if accurate, they should equal 0. Results were either 0 or less than 0.0001 above or below 0. The proportion of individuals with less than a high school education were added to individuals with at least a high school education with the expectation that they would equal 1. This was not the case: most totals ranged from 0.4 to 0.6. Referring to the documentation and exploring the data illuminated the reason. The Total population 15 years and over by highest level of schooling is a poorly defined population. The Census Profile data contains numerous population totals broken out by different characteristics. For example, the education variables include 'Total population 15 years and over by highest level of schooling', the marital status variables include 'Total population 15 years and over by marital status' and the labor force variables include 'Total population 15 years and over by labor force activity'. It was expected that summing together the number of individuals aged 15 years and over using the age breakouts in the Census Profile data would include the same population as these 'Total population 15 years and over by…' variables. This was checked for the 'Total population 15

years and over by highest level of schooling' variable. The difference between these two totals was more than can be explained by rounding error in the Census Profile data. Checking this variable against other variables that call themselves 'Total population 15 years and over by…' a sizeable and apparently unexplainable difference was found. Another drawback with the 2001 education aggregate variables is that in the 2001 Census Profile data education data is supplied for individuals aged 20 and over (Statistics Canada, 1999). By contrast, SLID education data was available for individuals aged 15 and over. This, added to the other problems described, provides the reason why the education aggregate variables do not match well.

Having identified an explanation for the low matching across education variables, similar explanations were sought for the employment variables. Three main limitations in both the Census Profiles and SLID documentation regarding ambiguous definitions of populations and variables were found that could explain the low number of matches across the employment variables. First, there was some ambiguity over the definition of the labor force. SLID defines the labor force as persons aged 16 to 69 who were employed during the survey reference period. Therefore, the employment variables used in the construction of the aggregate variables only refers to these individuals. The Census Profile data on the other hand defines the labor force as employed individuals aged 15 and over. Although this may cause some disparity, it is unlikely to be the only cause of the low number of matches. Further investigation revealed a more severe limitation regarding the classification of individual labor force status (Statistics Canada, 1997; Statistics Canada, 1999).

One of the strengths of the SLID as a longitudinal survey is that it asks for information on every job an individual has held during the reference year, rather than focusing on the job at the time of the survey. Regarding class of worker (paid worker, employee, self-employed, etc.) individuals can, therefore, hold several statuses, in addition, they are asked to report their status for each month so that they have 12 statuses over the year. By contrast, the Census Profile data only reports the class of worker for individuals at the time the Census is carried out. For the construction of the non_employee aggregate variable, it was necessary that individuals only fall into one class of worker category. However, the SLID uses the main job concept to categorize individuals into the class of worker variable clwrkr1. Main job is typically the job with the longest duration, greatest number of hours worked over the year, and most usual hours worked in a given month (Statistics Canada, 1997; Statistics Canada, 2007). Thus, the difference in the reference periods of the samples and the SLID's focus on main job is a possible explanation for lower matching rates.

The Census Profile data has its own ambiguities around the class of worker variable. In the Census Profile data on class of worker there is a category defined as Class of worker-Not Applicable (Statistics Canada, 1999). The documentation does not explain who this group consists of or what characteristics of individuals in this category make class of worker not applicable to them. In an effort to take this into account, the aggregate variable of non-employee was constructed using an all classes of worker variable as the denominator (a variable that does not include the class of worker-not applicable individuals). This was used in place of the variable total labor force 15 years and over by class of worker, which did include those individuals. Although this avoids using unclear population definitions as a denominator, it does not help explain where that category of individuals should be included most accurately in the class of worker categorization. These problems may explain why the non_employee variables are not matching well.

Finally, there was a difference in the definition of the Census Profile data and SLID managerial occupation variables that may render them incomparable. In the Census Profile data, individuals are asked to explain the type of job they have and their main responsibilities, and from this information they are coded into occupation classifications. This classification includes a section on management occupations that was used to produce the proportion of individuals with occupations perceived to be managerial variable (Statistics Canada, 1999). In the SLID, on the other hand, individuals are asked if they perceive their job to be managerial

(Statistics Canada, 1997). The self-identification involved here suggests that this variable is likely to be largely inconsistent: what defines a job as managerial is not clearly defined. Individuals may identify themselves as having a managerial job when in fact they do not. This could explain why the pct_mgt variable does not show a high number of matches.

Conclusion

The investigations and results are outlined here as a precaution to researchers wishing to create aggregate variables or use the SLID aggregate variables created in this study. The construction and comparison of aggregate variables should not be undertaken without caution. Despite their limitation, it is hoped that the constructed SLID aggregates could be of some use to researchers. The following points serve as a set of cautions to those wishing to use this approach based on difficulties that might be encountered in aggregate variable construction and comparison.

Internal Consistency

When constructing aggregate variables it is important that the variables are internally coherent. For example, imagine creating two aggregate education variables at the EA level; one is the proportion of individuals with less than a high school education, the other is the proportion of individuals with at least a high school education. If a researcher added the two proportions together across all EAs all totals should equal 1, if it does not, further investigation would be required to uncover reasons why.

Target Population

When creating and comparing aggregate variables it is important to know the target population of the variable. Some variables apply to individuals over a certain age, some apply to individuals who only answered positively to survey questions, and some apply to all respondents. This becomes more important if researchers wish to check the validity of their constructed aggregates by comparing the sample characteristics with the Census profile data characteristics. If constructing proportions of individuals with certain characteristics, it is important that the denominator be the same in both variables. The Census profile data documentation clearly defines its target population for each variable but it can be unclear how individuals were included. For example, in the 1996 Census profile data the 'Total population 15 years and over by highest level of schooling' was not the same as 'All individuals aged 15 years and over'. In some cases, the target populations for the same variable in the 1996 and 2001 Census profile data were different. For example, in the 1996 Census profile data, the education data was available for individuals aged 15 and over; and in the 2001 Census profile data, it was supplied for individuals aged 20 and over (Statistics Canada, 1999).

Definitions

It is important to understand how variables are defined in order to construct useful aggregate variables that are as accurate as possible. What a researcher may consider to be a standard classification may in fact be different across different datasets. In the example used in this article, it was found that the definition of labor force was not clear. In SLID, labor force is defined as persons aged 16 to 69 who were employed during the survey reference period. In the Census profile data, the labor force is defined as employed individuals aged 15 and over. This difference made comparing the Census profile and SLID aggregate employment variables inappropriate. Variable definitions may also have unexplained ambiguities that must be taken into account. For example, in the Census profile data there were ambiguities with the class of worker variable, which made comparison difficult.

Survey Design

The way in which surveys are designed can make constructing and comparing aggregate variables problematic. One of the strengths of the SLID as a longitudinal survey is that it asks for information on every job an individual has held during the reference year rather than focusing on the job at the time the survey is carried out. The result is that many records may exist for one individual. By contrast, the Census profile data holds one record per individual. Researchers must be clear on the survey design

and number of records per individual. If several records exist for each individual, some rationale must be used to select the most suitable record.

Classification

It is important to understand how variables have been coded into categories and whether individuals self-identify for certain classifications. This has implications for category definitions and how comparable they are across datasets. In the example provided, a difference was identified between the Census profile data and SLID in how managerial occupations are defined. The difference between coding by self-identification and coding by an external classifier is an important one that could lead to inconsistent definitions.

This article outlined the importance of aggregate level variables for use in multilevel analysis and introduced the idea of generating aggregate level variables in one dataset for use across other datasets. Generating and comparing aggregate variables was described using an example generating employment and education aggregate variables in the SLID for 1996 and 2001 cross-sectional samples at the ER Level and comparing them to similar estimates constructed using the 1996 and 2001 Census profile data. The difficulties encountered resulted in a set of cautions for researchers wishing to use this approach. As a whole, this article may serve as a guide to researchers in the generation and comparison of these or similar aggregate variables and also emphasizes the precautions that must be taken when using this approach.

References

Gonthier, D., Hotton, T., Cook, C., & Wilkins, R. (2006). Merging area-level census data with survey data in statistics Canada research data centres. *The Research Data Centres Information and Technical Bulletin*, *3*(*1*), 21-39.

Statistics Canada. 1997. Survey of labour and income dynamics: Microdata user's guide. *Dynamics of Labour and Income.* Catalogue No. 75M0001GPE. Ministry of Industry, Ottawa.

Statistics Canada. 1999. *1996 Census Dictionary, Final Edition Reference.* Catalogue No. 92-351-UIE. Ministry of Industry, Ottawa.

Statistics Canada. 2007. Guide to the Labour Force Survey. *Dynamics of Labour and Income.* Labour Statistics Division. Catalogue No. 71-543-GIE. Ministry of Industry, Ottawa.

Appendix

Tables A1 and A2 show the age and education verifications by ER for both the Census and the SLID for 1996 and 2001. The variable definitions are as follows:

SLIDage1:	'pct_15to25' + 'pct_25to49' + 'pct_50to74' + 'pct_75over'
SLIDage2:	'pct_15to25' + 'pct_25to49' + 'pct_50to64' + 'pct_65over'
SLIDage3:	Difference between 'pct_15to25' and ('pct_15to19' + 'pct_20to24')
SLIDage4:	SLIDage1 for females
SLIDage5:	SLIDage2 for females
SLIDage6:	SLIDage3 for females
Censage1:	'pct_15to25' + 'pct_25to49' + 'pct_50to74' + 'pct_75over'
Censage2:	'pct_15to25' + 'pct_25to49' + 'pct_50to64' + 'pct_65over'
Censage3:	Difference between 'pct_15to25' and ('pct_15to19' + 'pct_20to24')
Censage4:	Censusage1 for females
Censage5:	Censusage2 for females
Censage6:	Censusage3 for females
Censedu1:	'Less than high school' + 'at least high school'
Censedu2:	Difference between 'postsecondary certificate' and 'university certificate' + 'non-university certificate'

Table A1: 1996 Age and Education Verifications by ER for the Census and SLID

xerres25	SLIDage1	SLIDage2	SLIDage3	SLIDage4	SLIDage5	SLIDage6	Censage1	Censage2	Censage3	Censage4	Censage5	Censage6	Censedu1	Censedu2
1010	1	1	7.45E-09	1	1	0.00E+00	1.0001	1.0001	0	0.9990	0.9990	-7.45E-09	0.4942	-0.0006
1020	1	1	0	1	1	7.45E-09	1.0032	1.0032	0	1.0065	1.0065	0	0.7476	-0.0015
1030	1	1	7.45E-09	1	1	0	0.9990	0.9990	-7.45E-09	1.0020	1.0020	0	0.6251	-0.0002
1040	1	1	7.45E-09	1	1	7.45E-09	1.0001	1.0001	7.45E-09	1.0031	1.0031	7.45E-09	0.6711	-0.0010
1110	1	1	0	1	1	0.00E+00	0.9982	0.9982	7.45E-09	0.9957	0.9957	7.45E-09	0.5573	-0.0017
1210	1	1	0	1	1	0	0.9990	0.9990	0	0.9980	0.9980	-7.45E-09	0.5875	-0.0005
1220	1	1	0	1	1	0.00E+00	1.0028	1.0028	0	1.0051	1.0051	0	0.5713	-0.0023
1230	1	1	-7.45E-09	1	1	7.45E-09	1.0001	1.0001	-7.45E-09	0.9991	0.9991	0	0.5520	-0.0013
1240	1	1	0	1	1	-3.73E-09	1.0003	1.0003	-7.45E-09	0.9996	0.9996	-7.45E-09	0.6180	-0.0014
1250	1	1	-7.45E-09	1	1	7.45E-09	1.0002	1.0002	-7.45E-09	1.0001	1.0001	-7.45E-09	0.4052	-0.0018
1310	1	1	7.45E-09	1	1	0	1.0018	1.0018	0	1.0021	1.0021	-7.45E-09	0.6512	-0.0012
1320	1	1	0	1	1	-3.73E-09	0.9988	0.9988	7.45E-09	0.9998	0.9998	-7.45E-09	0.5567	-0.0019
1330	1	1	1.12E-08	1	1	0.00E+00	1.0014	1.0014	0	1.0009	1.0009	7.45E-09	0.5584	-0.0014
1340	1	1	0	1	1	0	1.0012	1.0012	-7.45E-09	1.0042	1.0042	0	0.5072	-0.0008
1350	0.9999999	1	7.45E-09	1	1	7.45E-09	0.9993	0.9993	-7.45E-09	1.0001	1.0001	7.45E-09	0.6550	-0.0018
2410	1	1	3.73E-09	1	1	0.00E+00	1.0004	1.0004	-7.45E-09	1.0035	1.0035	-7.45E-09	0.6955	-0.0021
2415	1	1	0	1	1	0	1.0030	1.0030	0	1.0033	1.0033	0	0.6180	-0.0018
2420	1	1	7.45E-09	1	1	-7.45E-09	0.9993	0.9993	0	0.9988	0.9988	-7.45E-09	0.4903	-0.0013
2425	1	1	-7.45E-09	0.9999999	0.9999999	3.73E-09	1.0004	1.0004	-7.45E-09	1.0009	1.0009	0	0.5995	-0.0018
2430	1	1	-7.45E-09	1	1	-3.73E-09	1.0011	1.0011	0	1.0032	1.0032	0	0.5755	-0.0013
2435	1	1	7.45E-09	1	1	-7.45E-09	0.9992	0.9992	7.45E-09	1.0004	1.0004	0	0.5359	-0.0016
2440	1	1	7.45E-09	1	1	7.45E-09	0.9998	0.9998	0	0.9997	0.9997	0	0.4741	-0.0016
2445	1	1	3.73E-09	1	1	7.45E-09	0.9998	0.9998	0	0.9997	0.9997	0	0.5175	-0.0005
2450	1	1	0	1	1	-1.86E-09	1.0013	1.0013	0	1.0021	1.0021	7.45E-09	0.6023	-0.0015
2455	1	1	7.45E-09	1	1	0	1.0003	1.0003	0	1.0005	1.0005	-7.45E-09	0.5817	-0.0019
2460	1	1	0	1	1	-3.73E-09	1.0012	1.0012	7.45E-09	1.0008	1.0008	-7.45E-09	0.5385	-0.0021
2465	1	1	-7.45E-09	1	1	-3.73E-09	1.0028	1.0028	0	1.0013	1.0013	0	0.6499	-0.0017
2470	1	1	0	1	1	0.00E+00	0.9986	0.9986	7.45E-09	0.9972	0.9972	-7.45E-09	0.5853	-0.0008
2475	1	1	7.45E-09	1	1	0.00E+00	1.0000	1.0000	-7.45E-09	1.0003	1.0003	-7.45E-09	0.5630	-0.0010
2480	1	1	-7.45E-09	1	1	3.73E-09	0.9986	0.9986	0	0.9954	0.9954	7.45E-09	0.6526	-0.0026
2490	1	1	0	1	1	7.45E-09	1.0008	1.0008	0	0.9978	0.9978	0	0.7966	-0.0015
3510	1	1	0	1	1	3.73E-09	0.9997	0.9997	0	0.9995	0.9995	0	0.4249	-0.0012
3520	1	1	-7.45E-09	1	1	0.00E+00	1.0008	1.0008	0	1.0016	1.0016	7.45E-09	0.5453	-0.0009
3530	1	1	7.45E-09	1	1	7.45E-09	0.9995	0.9995	7.45E-09	0.9995	0.9995	-7.45E-09	0.4462	-0.0009
3540	1	1	-7.45E-09	1	1	7.45E-09	0.9998	0.9998	-7.45E-09	1.0001	1.0001	0	0.5015	-0.0013
3550	1	1	7.45E-09	1	1	0	0.9999	0.9999	0	1.0000	1.0000	0	0.5109	-0.0011
4610	1	1	0	1	1	0.00E+00	1.0005	1.0005	0	0.9979	0.9979	0	0.6644	-0.0031
4620	1	1	-7.45E-09	1	1	0.00E+00	1.0031	1.0031	7.45E-09	1.0056	1.0056	0	0.7737	-0.0002
4630	1	1	-3.73E-09	1	1	-7.45E-09	0.9952	0.9952	0	0.9958	0.9958	0	0.6073	-0.0022
4640	1	1	-7.45E-09	1	1	-7.45E-09	0.9980	0.9980	7.45E-09	0.9921	0.9921	0	0.7333	-0.0010
4650	1	1	0	1	1	-3.73E-09	1.0012	1.0012	0	1.0009	1.0009	0	0.4713	-0.0014
4660	1	1	-7.45E-09	1	1	0	0.9990	0.9990	0	0.9986	0.9986	-7.45E-09	0.6325	-0.0020
4670	1	1	0	1	1	0	1.0026	1.0026	7.45E-09	1.0106	1.0106	0	0.8080	-0.0015
4680	1	1	0	1	1	0	1.0014	1.0014	0	1.0004	1.0004	-7.45E-09	0.7032	-0.0015
4710	1	1	7.45E-09	1	1	-7.45E-09	0.9979	0.9979	0	0.9980	0.9980	0	0.5170	-0.0006
4720	1	1	-7.45E-09	1	1	-3.73E-09	1.0012	1.0012	7.45E-09	1.0028	1.0028	0	0.6095	-0.0006
4730	1	1	0	1	1	0	1.0014	1.0014	0	1.0015	1.0015	0	0.4866	-0.0022
4740	1	1	3.73E-09	1	1	0	0.9975	0.9975	-3.73E-09	0.9965	0.9965	-3.73E-09	0.6997	-0.0003
4750	1	1	1.12E-08	1	1	3.73E-09	1.0025	1.0025	0	0.9970	0.9970	0	0.6341	-0.0029
4760	1	1	0	1	1	-1.49E-08	1.0029	1.0029	1.49E-08	1.0053	1.0053	0	0.8700	-0.0006
4810	1	1	7.45E-09	1	1	7.45E-09	0.9977	0.9977	0	0.9981	0.9981	0	0.5317	-0.0007
4820	1	1	-7.45E-09	1	1	-7.45E-09	1.0018	1.0018	0	1.0004	1.0004	-7.45E-09	0.6236	-0.0034
4830	1	1	-7.45E-09	1	1	0.00E+00	1.0000	1.0000	0	0.9996	0.9996	0	0.3959	-0.0013
4840	1	1	0	1	1	3.73E-09	1.0022	1.0022	-7.45E-09	1.0037	1.0037	0	0.5745	-0.0022
4850	1	1	7.45E-09	1	1	0.00E+00	0.9987	0.9987	-7.45E-09	0.9993	0.9993	0	0.5427	-0.0025
4860	1	1	-7.45E-09	1	1	0.00E+00	0.9995	0.9995	0	1.0001	1.0001	0	0.4481	-0.0012
4870	1	1	7.45E-09	1	1	0	0.9994	0.9994	0	0.9985	0.9985	-7.45E-09	0.5980	-0.0010
4880	1	1	0	1	1	0.00E+00	1.0016	1.0016	7.45E-09	1.0037	1.0037	0	0.5616	-0.0008
5910	1	1	7.45E-09	1	1	7.45E-09	0.9998	0.9998	0	0.9990	0.9990	0	0.4349	-0.0020
5920	1	1	-7.45E-09	1	1	0.00E+00	1.0007	1.0007	-7.45E-09	1.0005	1.0005	-7.45E-09	0.4168	-0.0015
5930	1	1	0	1	1	0.00E+00	0.9990	0.9990	0	0.9978	0.9978	0	0.5033	-0.0016
5940	0.9999999	1	3.73E-09	1	1	0	1.0024	1.0024	7.45E-09	1.0020	1.0020	7.45E-09	0.5081	-0.0029
5950	1	1	-3.73E-09	1	1	1.12E-08	1.0026	1.0026	7.45E-09	1.0045	1.0045	-7.45E-09	0.5574	-0.0013
5960	1	1	0	1	1	-7.45E-09	0.9976	0.9976	0	0.9977	0.9977	0	0.6184	-0.0019
5970	1	1	0	1	1	0	0.9998	0.9998	0	1.0059	1.0059	7.45E-09	0.6671	-0.0030
5980	1	1	-7.45E-09	1	1	-7.45E-09	0.9946	0.9946	-7.45E-09	0.9922	0.9922	0	0.6399	-0.0007

Table A2: 2001 Age and Education Verifications by ER for the Census and SLID

erres25	SLIDage1	SLIDage2	SLIDage3	SLIDage4	SLIDage5	SLIDage6	Censage1	Censage2	Censage3	Censage4	Censage5	Censage6	Censedu1	Censedu2
1010	1	1	0	1	1	0	0.9986	0.9986	7.45E-09	0.9987	0.9987	0	0.4257	-0.0013
1020	1	1	7.45E-09	1	1	-3.73E-09	1.0019	1.0019	-7.45E-09	1.0033	1.0033	7.45E-09	0.7350	-0.0023
1030	1	1	0	1	1	-7.45E-09	1.0012	1.0012	-7.45E-09	1.0008	1.0008	0	0.5638	-0.0041
1040	1	1	-3.73E-09	1	1	0	1.0001	1.0001	0	0.9990	0.9990	7.45E-09	0.6387	-0.0041
1110	1	1	0	1	1	-7.45E-09	0.9964	0.9964	-7.45E-09	0.9944	0.9944	0	0.4963	-0.0029
1210	1	1	0	1	1	-7.45E-09	1.0010	1.0010	7.45E-09	1.0017	1.0017	0	0.5270	-0.0019
1220	1	1	3.73E-09	1	1	3.73E-09	0.9989	0.9989	0	0.9997	0.9997	0	0.5026	-0.0026
1230	1	1	0	1	1	3.73E-09	0.9990	0.9990	-7.45E-09	0.9992	0.9992	3.73E-09	0.4937	-0.0030
1240	1	1	-3.73E-09	1	1	-3.73E-09	1.0010	1.0010	-3.73E-09	0.9975	0.9975	3.73E-09	0.5593	-0.0008
1250	1	1	0	1	1	-3.73E-09	0.9998	0.9998	-7.45E-09	0.9999	0.9999	7.45E-09	0.3342	-0.0020
1310	1	1	7.45E-09	1	1	-7.45E-09	0.9980	0.9980	0	0.9994	0.9994	7.45E-09	0.6102	-0.0019
1320	1	1	-7.45E-09	1	1	7.45E-09	1.0002	1.0002	0	0.9988	0.9988	0	0.4938	-0.0011
1330	1	1	0	1	1	-3.73E-09	0.9959	0.9959	0	0.9943	0.9943	-7.45E-09	0.4994	-0.0030
1340	1	1	0	1	1	0	0.9991	0.9991	0	0.9991	0.9991	7.45E-09	0.4530	-0.0017
1350	1	1	-7.45E-09	1	1	7.45E-09	0.9995	0.9995	0	0.9993	0.9993	7.45E-09	0.6180	-0.0041
2410	1	1	-3.73E-09	1	1	3.73E-09	1.0040	1.0040	0	1.0022	1.0022	-7.45E-09	0.6759	-0.0058
2415	1	1	0	1	1	0	0.9995	0.9995	0	0.9990	0.9990	0	0.5699	-0.0113
2420	1	1	0	1	1	0	1.0008	1.0008	7.45E-09	0.9996	0.9996	7.45E-09	0.4307	-0.0074
2425	1	1	0	1	1	0	0.9996	0.9996	0	0.9995	0.9995	0	0.5394	-0.0091
2430	1	1	0	1	1	0	0.9989	0.9989	7.45E-09	0.9971	0.9971	7.45E-09	0.5173	-0.0084
2433	1	1	3.73E-09	1	1	-1.12E-08	1.0006	1.0006	0	1.0014	1.0014	0	0.5702	-0.0096
2435	1	1	-7.45E-09	1	1	-3.73E-09	1.0005	1.0005	0	1.0000	1.0000	0	0.4800	-0.0074
2440	1	1	-3.73E-09	1	1	3.73E-09	1.0003	1.0003	0	1.0008	1.0008	3.73E-09	0.4058	-0.0063
2445	1	1	1.86E-09	1	1	1.86E-09	0.9996	0.9996	0	1.0001	1.0001	0	0.4621	-0.0068
2450	1	1	3.73E-09	1	1	0	1.0000	1.0000	7.45E-09	1.0023	1.0023	7.45E-09	0.5522	-0.0071
2455	1	1	7.45E-09	1	1	0	0.9993	0.9993	7.45E-09	1.0004	1.0004	7.45E-09	0.5127	-0.0063
2460	1	1	0	1	1	0	1.0002	1.0002	-7.45E-09	0.9992	0.9992	0	0.4808	-0.0077
2465	1	1	0	1	1	7.45E-09	1.0012	1.0012	-7.45E-09	1.0003	1.0003	0	0.6077	-0.0075
2470	1	1	0	1	1	-7.45E-09	0.9996	0.9996	7.45E-09	0.9996	0.9996	-7.45E-09	0.5316	-0.0075
2475	1	1	-7.45E-09	1	1	-7.45E-09	0.9990	0.9990	7.45E-09	0.9994	0.9994	7.45E-09	0.5064	-0.0113
2480	1	1	0	1	1	-7.45E-09	0.9983	0.9983	-7.45E-09	0.9991	0.9991	0	0.6162	-0.0092
2490	1	1	0	1	1	0	0.9897	0.9897	-7.45E-09	0.9902	0.9902	7.45E-09	0.7877	-0.0080
3510	1	1	3.73E-09	1	1	3.73E-09	0.9999	0.9999	-7.45E-09	0.9995	0.9995	-7.45E-09	0.3395	-0.0011
3515	1	1	0	1	1	0	1.0015	1.0015	7.45E-09	1.0021	1.0021	0	0.4547	-0.0014
3520	1	1	-7.45E-09	1	1	-7.45E-09	1.0014	1.0014	0	1.0011	1.0011	7.45E-09	0.4763	-0.0015
3530	1	1	0	1	1	-3.73E-09	1.0000	1.0000	7.45E-09	1.0002	1.0002	0	0.3668	-0.0015
3540	1	1	0	1	1	0	0.9996	0.9996	7.45E-09	0.9985	0.9985	0	0.4263	-0.0011
3550	1	1	0	1	1	7.45E-09	0.9996	0.9996	0	1.0003	1.0003	0	0.4439	-0.0012
3560	1	1	-7.45E-09	1	1	0	1.0012	1.0012	-7.45E-09	0.9997	0.9997	7.45E-09	0.4281	-0.0018
3570	1	1	7.45E-09	1	1	-3.73E-09	1.0000	1.0000	-7.45E-09	1.0006	1.0006	0	0.4608	-0.0014
3580	1	1	0	1	1	0	1.0020	1.0020	0	1.0019	1.0019	-7.45E-09	0.5108	-0.0018
3590	1	1	0	1	1	0	0.9997	0.9997	7.45E-09	1.0005	1.0005	-7.45E-09	0.4850	-0.0024
3595	1	1	0	1	1	-7.45E-09	1.0025	1.0025	7.45E-09	1.0011	1.0011	0	0.4838	-0.0023
4610	1	1	7.45E-09	1	1	3.73E-09	0.9972	0.9972	7.45E-09	0.9947	0.9947	-7.45E-09	0.6085	-0.0028
4620	1	1	-7.45E-09	1	1	0	0.9950	0.9950	0	0.9928	0.9928	0	0.7222	-0.0042
4630	1	1	-7.45E-09	1	1	7.45E-09	0.9986	0.9986	0	0.9936	0.9936	-7.45E-09	0.5573	-0.0010
4640	1	1	3.73E-09	1	1	7.45E-09	1.0028	1.0028	-7.45E-09	1.0017	1.0017	7.45E-09	0.6848	-0.0010
4650	1	1	-7.45E-09	1	1	-3.73E-09	1.0002	1.0002	0	1.0003	1.0003	0	0.4066	-0.0022
4660	1	1	-7.45E-09	1	1	-3.73E-09	1.0005	1.0005	0	0.9978	0.9978	-3.73E-09	0.5578	-0.0012
4670	1	1	3.73E-09	1	1	3.73E-09	0.9977	0.9977	0	0.9983	0.9983	-7.45E-09	0.7399	0.0005
4680	1	1	0	1	1	7.45E-09	1.0027	1.0027	7.45E-09	1.0017	1.0017	0	0.6551	-0.0015
4710	1	1	-7.45E-09	1	1	7.45E-09	1.0015	1.0015	0	1.0003	1.0003	-7.45E-09	0.4402	-0.0014
4720	1	1	0	1	1	-7.45E-09	1.0006	1.0006	0	0.9972	0.9972	7.45E-09	0.5575	-0.0012
4730	1	1	0	1	1	-7.45E-09	0.9996	0.9996	0	0.9985	0.9985	0	0.4101	-0.0029
4740	1	1	-3.73E-09	1	1	0	0.9976	0.9976	0	0.9989	0.9989	-7.45E-09	0.6562	-0.0028
4750	1	1	7.45E-09	1	1	-7.45E-09	0.9999	0.9999	0	1.0013	1.0013	0	0.5572	-0.0033
4760	1	1	3.73E-09	1	1	-3.73E-09	0.9990	0.9990	-1.49E-08	1.0050	1.0050	-7.45E-09	0.8202	-0.0024
4810	1	1	0	1	1	7.45E-09	0.9998	0.9998	0	1.0004	1.0004	0	0.4730	-0.0012
4820	1	1	-7.45E-09	1	1	7.45E-09	0.9970	0.9970	-7.45E-09	0.9999	0.9999	-7.45E-09	0.5297	-0.0029
4830	1	1	0	1	1	-7.45E-09	0.9995	0.9995	7.45E-09	0.9990	0.9990	0	0.3180	-0.0022
4840	1	1	7.45E-09	1	1	0	1.0009	1.0009	7.45E-09	1.0021	1.0021	0	0.4995	-0.0038
4850	1	1	0	1	1	-7.45E-09	1.0010	1.0010	0	0.9986	0.9986	7.45E-09	0.4609	-0.0029
4860	1	1	-7.45E-09	1	1	-7.45E-09	0.9996	0.9996	7.45E-09	0.9994	0.9994	0	0.3719	-0.0023
4870	1	1	7.45E-09	1	1	0	0.9994	0.9994	7.45E-09	0.9953	0.9953	-7.45E-09	0.5274	-0.0029
4880	1	1	-7.45E-09	1	1	7.45E-09	0.9996	0.9996	-7.45E-09	1.0017	1.0017	7.45E-09	0.4807	-0.0030
5910	1	1	3.73E-09	1	1	0	1.0011	1.0011	-7.45E-09	1.0014	1.0014	0	0.3641	-0.0025
5920	1	1	3.73E-09	1	1	-7.45E-09	1.0003	1.0003	0	1.0008	1.0008	0	0.3461	-0.0026
5930	1	1	0	1	1	-7.45E-09	1.0007	1.0007	0	1.0019	1.0019	0	0.4322	-0.0022
5940	1	1	7.45E-09	1	1	3.73E-09	0.9987	0.9987	3.73E-09	0.9984	0.9984	-7.45E-09	0.4417	-0.0035
5950	1	1	0	1	1	-3.73E-09	0.9987	0.9987	0	0.9980	0.9980	0	0.4733	-0.0031
5960	1	1	-3.73E-09	1	1	-7.45E-09	1.0010	1.0010	7.45E-09	1.0028	1.0028	0	0.5590	-0.0034
5970	1	1	-7.45E-09	1	1	0	0.9976	0.9976	0	0.9980	0.9980	0	0.6221	-0.0075
5980	1	1	7.45E-09	1	1	-7.45E-09	0.9955	0.9955	7.45E-09	0.9946	0.9946	7.45E-09	0.5680	-0.0005

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 626-631

1538 – 9472/09/$95.00

Markov Modeling of Breast Cancer

Chunling Cong Chris P. Tsokos
University of South Florida

Previous work with respect to the treatments and relapse time for breast cancer patients is extended by applying a Markov chain to model three different types of breast cancer patients: alive without ever having relapse, alive with relapse, and deceased. It is shown that combined treatment of tamoxifen and radiation is more effective than single treatment of tamoxifen in preventing the recurrence of breast cancer. However, if the patient has already relapsed from breast cancer, single treatment of tamoxifen would be more appropriate with respect to survival time after relapse. Transition probabilities between three stages during different time periods, 2-year, 4-year, 5-year, and 10-year, are also calculated to provide information on how likely one stage moves to another stage within a specific time period.

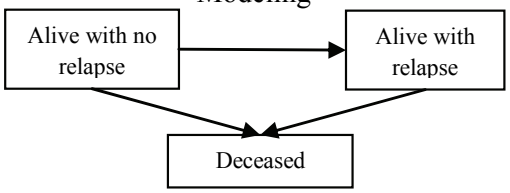
Key words: Markov chain, breast cancer, relapse time, tamoxifen and radiation.

Introduction

The Markov (1906) chain model has been applied in various fields such as physics, queuing theory, internet application, economics, finance, and social sciences among others. As an effective and efficient way of describing a process in which an individual moves through a series of states (stages) in continuous time, homogeneous Markov models have also been extensively used in health sciences where the progression of certain diseases are of great importance to both doctors and patients. In the present study, the main objective is to investigate the progression of breast cancer in patients in three different stages who were given different treatments. One group of patients received combined treatments of tamoxifen and radiation, and the other group received only tamoxifen. Figure 1 shows the three stages of interest in the study are: alive with no relapse, alive with relapse, and deceased. Even though breast cancer patients who have recurrence may be treated and recover from breast cancer to become active with no relapse, due to the fact that the data does not include any observations of that process, we consider the second state-alive with relapse as those patients who once had relapse and are still alive, regardless of whether they have recovered from breast cancer or not.

Chunling Cong is a doctoral student in Statistics at the University of South Florida. Her research interests are in developing forecasting and statistical analysis and modeling of cancer. Email: ccong@mail.usf.edu. Chris Tsokos is a Distinguished University Professor in mathematics and Statistics at the University of South Florida. His research interests are in modeling Global Warming, analysis and modeling of cancer data, parametric, Bayesian and nonparametric reliability, and stochastic systems, among others. He is a fellow of both ASA and ISI. Email: profcpt@cas.usf.edu.

Figure 1: Three Stages of Breast Cancer Modeling

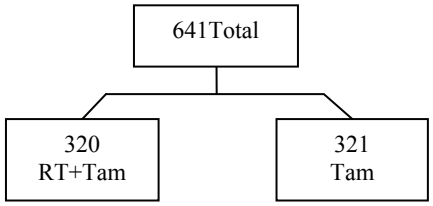


Methodology

Between December 1992 and June 2000, a total of 769 women were enrolled and randomized in the study. Among these, 386 received combined radiation and tamoxifen (RT+Tam), and the

remaining 383 received tamoxifen (Tam) only. The last follow-up was conducted in the summer of 2002. As shown in Figure 2, only those 641 patients enrolled at the Princess Margaret Hospital are included: 320 and 321 in RT+Tam and Tam treatment groups, respectively.

Figure 2: Breast Cancer Data



This data was used by Fyles, et al. and was later analyzed by Ibrahim, et al. Analysis was conducted on this data with respect to the treatment effect of the two different treatments using decision tree and modeled relapse time using AFT and Cox-PH model. Mixture models were also applied to compare the cure rate of the two groups.

The Markov Chain Model

The Markov chain is a model for a finite or infinite random process sequence $X = \{X_1, X_{2,...,X_N}\}$. Unlike the independent identical distribution (i.i.d) model that assumes the independency of a sequence of events X_i's, the Markov model takes into account the dependencies among the X_i's.

Consider a random process $X = \{X_t\}_{t\geq 1} = \{X_1, X_2, ...\}$ of random variables taking values in a discrete set space of stages $S = \{1, 2, 3, ..., s\}$ where X_t represents the state of the process of an individual at time *t*. The transitions possible among the three stages in this study, alive without relapse, alive with relapse, and deceased are shown in Figure 1 indicated by arrows. Consider a realization of the history of the process up to and including time *t*, as $\{X_t = x_t, X_{t-1} = x_{t-1}, ..., X_1 = x_1\}$, where $x_t, x_{t-1}, ..., x_1$ is a sequence of stages at different times. A random process is called a Markov Chain if the conditional probabilities between the stages at different times satisfy the Markov property: the conditional probability of future one-step-event conditioned on the entire past of the process is just conditioned on the present stage of the process. In other words, the one-step future stage depends only on the present stage:

$$P(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, ..., X_1 = x_1) \quad (1)$$
$$= P(X_{t+1} = x_{t+1} \mid X_t = x_t)$$

for every sequence $x_1, ..., x_t, x_{t+1}$ of elements of S and every $t \geq 1$.

The transition probability from stage i to stage j at time t and transition intensity are defined by

$$p_{ij}(t) = p(X_{t+1} = j \mid X_t = i), \quad (2)$$

and

$$q_{ij}(t) = \lim_{h\to 0} \frac{P(X(t+h) = j \mid X(t) = i)}{h}, \quad (3)$$

where h is the time interval.

If the transition probabilities do not depend on time, $p_{ij}(t)$ can simply be written as p_{ij}, then the Markov chain is called time-homogeneous. If not specified, the following analysis is based on time-homogeneous Markov chain. A transition probability matrix $P(t)$ consisting of all the transition probabilities between stages in a matrix form is given by:

$$P(t) = \begin{Bmatrix} p_{11}(t) & p_{12}(t) & ... & p_{1s}(t) \\ p_{21}(t) & p_{22}(t) & ... & p_{2s}(t) \\ ... & ... & ... & ... \\ p_{s1}(t) & p_{s2}(t) & ... & p_{ss}(t) \end{Bmatrix}, \quad (4)$$

where probabilities in each row add up to 1. Thus, it is 100% certain that for any individual at time *t* is in one of the stages and the sum of probabilities of being in each stage is 1.

The transition probability matrix can be calculated by taking the matrix exponential of the scaled transition intensity matrix defined by

$$P(t) = Exp(tQ), \quad (5)$$

where

$$Q = \begin{Bmatrix} q_{11} & q_{12} & \cdots & q_{1s} \\ q_{21} & q_{22} & \cdots & q_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ q_{s1} & q_{s2} & \cdots & q_{ss} \end{Bmatrix}, \quad (6)$$

and q_{ij} denotes the transition intensity from stage i to stage j.

The exponential of a matrix A is defined by

$$Exp(A) = 1 + A^2/2! + A^3/3! + \ldots, \quad (7)$$

where each summand in the series is the matrix products. In this manner, once the intensity matrix is given, the transition probabilities can be calculated as shown above.

Next, the intensity matrix and transition probabilities matrix can be obtained by maximizing the likelihood $L(Q)$ which is a function of Q. Consider an individual consisting of a series of times $(t_1, t_2, ..., t_n)$ and corresponding stages $(x_1, x_2, ..., x_n)$. More specifically, consider a pair of successive stages observed to be i and j at time t_i and t_j. Three scenarios are proposed and considered here.

Scenario 1

If the information for the individual is obtained at arbitrary observation times (the exact time of the transition of stages is unknown) the contribution to the likelihood from this pair of states is:

$$L_{ij} = p_{ij}(t_j - t_i). \quad (8)$$

Scenario 2

If the exact times of transitions between different stages are recorded and there is no transition between the observation times, the contribution to the likelihood from this pair of stages is:

$$L_{ij} = p_{ij}(t_j - t_i) q_{ij}. \quad (9)$$

Scenario 3

If the time of death is known or $j = death$, but the stage on the previous instant before death is unknown as denoted by k (k could be any possible stage between stage i and death), the contribution to the likelihood function from this pair of stages is:

$$L_{ij} = \sum_{k \neq j} p_{ik}(t_j - t_i) q_{kj}. \quad (10)$$

Results

The breast cancer patients were divided into two groups RT+Tam and Tam based on the different treatments they received. For those patients who received combined treatments, 26 patients experienced relapse, 13 patients died without recurrence of breast cancer during the entire period of the study, and 14 died after recurrence of breast cancer. For the patients in the Tam group, 51 patients experienced relapse, 10 died without reoccurrence of breast cancer, and 13 died after recurrence of breast cancer.

As can be observed from the transition intensity matrixes for both groups RT+Tam and Tam as shown in Tables 1 and 2, patients who received single treatment have a higher transition intensity form Stage 1 to Stage 2, thus, they are more likely to have breast cancer recurrence. Thus, the probability of that happening in the Tam group is higher than that of the RT+Tam group. For those patients who died without relapse, there is no significant difference between the two treatments as illustrated by the intensity form Stage 1 to Stage 3.

Combined treatment is also more effective than a single treatment with respect to the possibility of death without relapse as can be observed from the transition intensity from Stage 1 to Stage 3. However, for those who already experienced relapse of breast cancer, patients who received combined treatments are more likely to die than those who received a single treatment. Therefore, combined treatment should be chosen over single treatment to avoid recurrence, but for those patients who already had breast cancer relapse, it would be advisable to choose a single treatment to extend the time from recurrence to death.

Figures 3 and 4 illustrate the effectiveness of the two treatments with respect to the survival probabilities and also show the survival curves of the patients who had recurrence and who had no recurrence in each treatment group.

From the above analysis, the proposed Markov chain model provides recommendations for which treatment to choose for breast cancer patients with respect to relapse and survival time. Moreover, it provides patients with very important information on the exact time or possibilities of recurrence and death. Estimated mean sojourn times in each transient stage for patients who received combined treatment are 43.46 and 3.25 in Stage 1 and Stage 2, respectively. Estimated mean sojourn times for patients who received single treatment are 25.53 and 11.72 in Stage 1 and Stage 2. This further confirms that patients with combined treatment will remain in Stage 1 longer than those with single treatment; however, for patients who had relapse of breast cancer, patients with single treatment will stay alive longer than those with combined treatment.

Another goal of this study was to provide a transition probability matrix at different times so that given a specific time period, the probability that a patient in a given stage will transit to another stage could be conveyed. Tables 5a-8b give 2-year, 4-year, 5-year and 10-year transition probability matrixes of patients in RT+Tam and Tam.

Conclusion

Through Markov chain modeling of the three stages of breast cancer patients , it has been shown that combined treatment of tamoxifen and radiation is more effective than single treatment of tamoxifen in preventing the recurrence of breast cancer. However, for patients who had a relapse of breast cancer, single treatment of tamoxifen proves to be more effective than combined treatment with respect to the survival probability. This finding could give significant guidance to doctors with respect to which breast cancer treatment should be given to breast cancer patients in different stages. Transition probabilities between different stages during 2 years, 4 years, 5 years and 10 years are also calculated for predicting purposes. Those transition probabilities could help provide a clearer view of how one stage transits to another stage within a given time period.

Acknowledgements

We wish to thank N.A Ibrahim for supplying us the source of the data that made the subject study possible. We also wish to express appreciation to Dr. James Kepner, Vice President of the American Cancer Society for our useful discussions on the present study.

References

Dynkin, E. B. (2006). *Theory of Markov Processes*. Dover Publications.

Fyles, A. W., & McCready, D. R. (2004). Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer, *New England Journal of Medicine*, *351*, 963-970.

Gentleman, R. C., Lawless J. F., Lindsey J. C., & Yan, P. (1994). Multi-State Markov models for analyzing incomplete disease history data with illustrations for HIV disease. *Statist. Med*., *13*, 805-821.

Ibrahim N. A., et al. (2008). Decision tree for competing risks survival probability in breast cancer study, *International Journal of Biomedical Sciences*, Volume 3 Number 1.

Jackson, C. (2007). *Multi-State modeling with R: The msm package, version 0.7.4.1 October*.

Kalblfleisch, J. D., & Lawless J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, *80*, 863-871.

Kay, R. (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics*, *42*, 855-865.

Lu, Y. & Stitt, F. W. (1994). Using Markov processes to describe the prognosis of HIV-1 infection. *Medical Decision Making*, *14*, 266-272.

Markov, A. (1906). Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, tom 15, *9 4*, 135-156.

Norris J. R. (1998). *Markov Chains*, Cambridge University Press.

Satten, G. A. & Longini, L. M. (1996). Markov chains with measurement error: estimating the true course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistician*, *45*, 275-309.

Sharples, L. D. (1993). Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. *Statistics in Medicine*, *12*, 1155-1169.

Table 1: Transition Intensity Matrix of RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	-0.02301	0.01957	0.0034
Stage 2	0	-0.3074	0.3074
Stage 3	0	0	0

Table 2: Transition Intensity Matrix of Tam

	Stage 1	Stage 2	Stage 3
Stage 1	-0.03917	0.03528	0.003889
Stage 2	0	-0.08533	0.08533
Stage 3	0	0	0

Figure 3: Survival Curves of Patients in RT+Tam

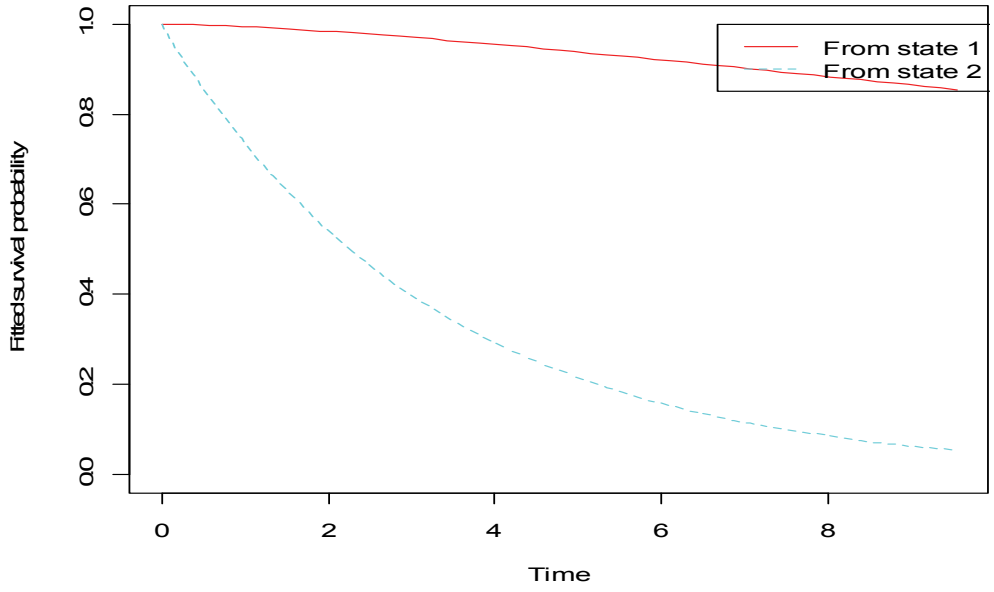


Figure 4: Survival Curves of Patients in Tam

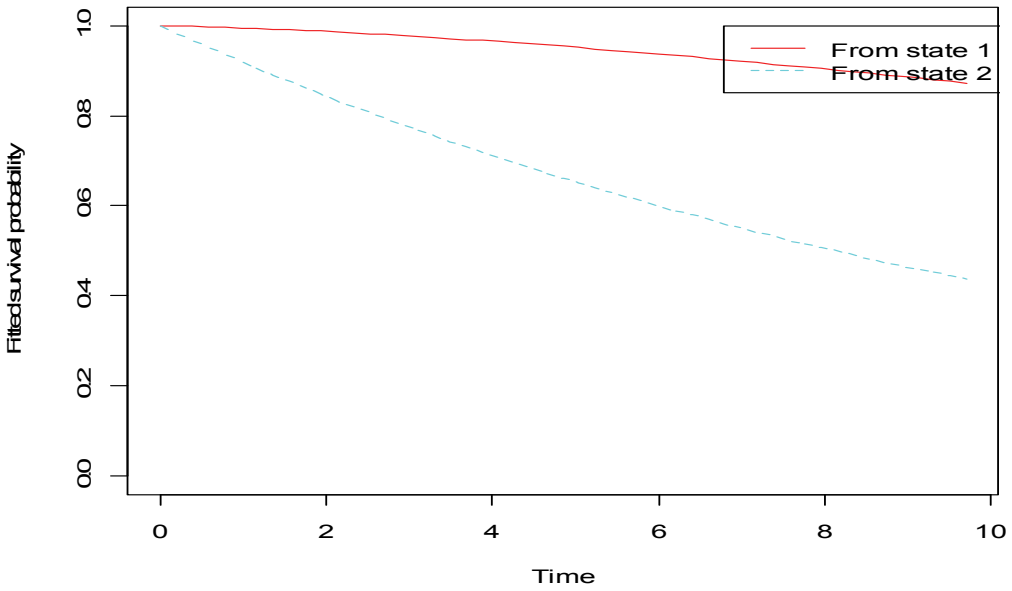


Table 5a: 2-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9550	0.0285	0.0165
Stage 2	0	0.5408	0.4592
Stage 3	0	0	0

Table 5b: 2-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9247	0.0623	0.0130
Stage 2	0	0.8431	0.1569
Stage 3	0	0	0

Table 6a: 4-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9121	0.0426	0.0453
Stage 2	0	0.2925	0.7075
Stage 3	0	0	0

Table 6b: 4-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8550	0.1102	0.0348
Stage 2	0	0.7108	0.2892
Stage 3	0	0	0

Table 7a: 5-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8913	0.0466	0.0621
Stage 2	0	0.2151	0.7849
Stage 3	0	0	0

Table 7b: 5-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8221	0.1295	0.0484
Stage 2	0	0.6527	0.3473
Stage 3	0	0	0

Table 8a: 10-year transition matrix for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.7945	0.0515	0.1540
Stage 2	0	0.0463	0.9537
Stage 3	0	0	0

Table 8b: 10-year transition matrix for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.6759	0.1910	0.1331
Stage 2	0	0.4260	0.5740
Stage 3	0	0	0

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 632-645

1538 – 9472/09/$95.00

STATISTICAL SOFTWARE APPLICATIONS & REVIEW
Ordinal Regression Analysis: Fitting the Proportional Odds Model Using Stata, SAS and SPSS

Xing Liu
Eastern Connecticut State University

Researchers have a variety of options when choosing statistical software packages that can perform ordinal logistic regression analyses. However, statistical software, such as Stata, SAS, and SPSS, may use different techniques to estimate the parameters. The purpose of this article is to (1) illustrate the use of Stata, SAS and SPSS to fit proportional odds models using educational data; and (2) compare the features and results for fitting the proportional odds model using Stata OLOGIT, SAS PROC LOGISTIC (ascending and descending), and SPSS PLUM. The assumption of the proportional odds was tested, and the results of the fitted models were interpreted.

Key words: Proportional Odds Models, Ordinal logistic regression, Stata, SAS, SPSS, Comparison.

Introduction

The proportional odds (PO) model, also called cumulative odds model (Agresti, 1996, 2002; Armstrong & Sloan, 1989; Long, 1997, Long & Freese, 2006; McCullagh, 1980; McCullagh & Nelder, 1989; Powers & Xie, 2000; O'Connell, 2006), is a commonly used model for the analysis of ordinal categorical data and comes from the class of generalized linear models. It is a generalization of a binary logistic regression model when the response variable has more than two ordinal categories. The proportional odds model is used to estimate the odds of being at or below a particular level of the response variable. For example, if there are j levels of ordinal outcomes, the model makes J-1 predictions, each estimating the cumulative probabilities at or below the j^{th} level of the outcome variable. This model can estimate the odds of being at or beyond a particular level of the response variable as well, because below and beyond a particular category are just two complementary directions.

Researchers currently have a variety of options when choosing statistical software packages that can perform ordinal logistic regression models. For example, some general purpose statistical packages, such as Stata, SAS and SPSS, all provide the options of analyzing proportional odds models. However, these statistical packages may use different techniques to estimate the ordinal logistic models. Long and Freese (2006) noted that Stata estimates cut-points in the ordinal logistic model while setting the intercept to be 0; other statistical software packages might estimate intercepts rather than cut-points. Agresti (2002) introduced both the proportional odds model and the latent variable model, and stated that parameterization in SAS (Proc Logistic) followed the formulation of the proportional odds model rather than the latent variable model. Hosmer and Lemeshow (2000) used a formulation which was consistent with Stata's expression to define the ordinal regression model by negating the logit coefficients.

Because statistical packages may estimate parameters in the ordinal regression model differently following different equations, the outputs they produce may not be the same, and thus they seem confusing to applied

Xing Liu is an Assistant Professor at Eastern Connecticut State University. Email: liux@easternct.edu. This article was presented at the 2007 Annual Conference of the American Educational Research Association (AERA) in Chicago, IL.

statisticians and researchers. Researchers are more likely to make mistakes in interpreting the results if ignoring the differences in parameter estimations using different software packages.

It is the aim of the article to clarify the misunderstanding and confusion when fitting ordinal regression models. To date, no study has been conducted to demonstrate fitting the proportional odds model using three general-purpose statistical packages, comparing differences and identifying similarities among them. Thus, this article seeks to fill this gap by: (1) demonstrating the use of Stata, SAS and SPSS to fit the proportional odds model; and (2) comparing the features and results for fitting the proportional odds model using Stata OLOGIT, SAS PROC LOGISTIC (ascending and descending), and SPSS PLUM. Data from a survey instrument TPGP (Teachers' Perceptions of Grading Practices) is used to demonstrate the PO analysis.

Theoretical Framework

In an ordinal logistic regression model, the outcome variable is ordered, and has more than two levels. For example, students' SES is ordered from low to high; childrens' proficiency in early reading is scored from level 0 to 5; and a response scale of a survey instrument is ordered from strongly disagree to strongly agree. One appealing way of creating the ordinal variable is via categorization of an underlying continuous variable (Hosmer & Lemeshow, 2000).

In this article, the ordinal outcome variable is teachers' teaching experience level, which is coded as 1, 2, or 3 (1 = low; 2 = medium; and 3 = high) and is categorized based on a continuous variable, teaching years. Teachers with less than five years of experience are categorized in the low teaching experience level; those with between 6 and 15 years are categorized in the medium level; and teachers with 15 years or more are categorized in the high level. The distribution of teaching years is highly positively skewed. The violation of the assumption of normality makes the use of Multiple Regression inappropriate. Therefore, the ordinal logistic regression is the most appropriate model for analyzing the ordinal outcome variable in this case.

A Latent-Variable Model

The ordinal logistic regression model can be expressed as a latent variable model (Agresti, 2002; Greene, 2003; Long, 1997, Long & Freese, 2006; Powers & Xie, 2000; Wooldridge & Jeffrey, 2001). Assuming a latent variable, Y* exists, $Y^* = \mathbf{x\beta} + \varepsilon$, can be defined where $\mathbf{x}$ is a row vector (1* k) containing no constant, $\boldsymbol{\beta}$ is a column vector (k*1) of structural coefficients, and ε is random error with standard normal distribution: $\varepsilon \sim N(0, 1)$.

Let Y* be divided by some cut points (thresholds): $\alpha_1, \alpha_2, \alpha_3 \ldots \alpha_j$, and $\alpha_1<\alpha_2<\alpha_3 \ldots < \alpha_j$. Considering the observed teaching experience level is the ordinal outcome, y, ranging from 1 to 3, where 1= low, 2 = medium and 3 = high, define:

$$Y = \begin{cases} 1 & if\ y^* \le \alpha_1 \\ 2 & if\ \alpha_1 < y^* \le \alpha_2 \\ 3 & if\ \alpha_2 < y^* \le \infty \end{cases}$$

Therefore, the probability of a teacher at each experience level can be computed. For example,

$$\begin{aligned} P(y = 1) &= P(y^* \le \alpha_1) \\ &= P(\mathbf{x\beta} + \varepsilon \le \alpha_1) \\ &= F(\alpha_1 - \mathbf{x\beta}); \end{aligned}$$

$$\begin{aligned} P(y = 2) &= P(\alpha_1 < y^* \le \alpha_2) \\ &= F(\alpha_2 - \mathbf{x\beta}) - F(\alpha_1 - \mathbf{x\beta}); \end{aligned}$$

$$\begin{aligned} P(y = 3) &= P(\alpha_2 < y^* \le \infty) \\ &= 1 - F(\alpha_2 - \mathbf{x\beta}); \end{aligned}$$

The cumulative probabilities can also be computed using the form:

$$P(Y \le j) = F(\alpha_j - \mathbf{x\beta}), \text{ where } j = 1, 2, \ldots J\text{-}1. \quad (1)$$

General Logistic Regression Model

In a binary logistic regression model, the response variable has two levels, with 1 = success of the events, and 0 = failure of the events. The probability of success is predicted on a set of predictors. The logistic regression model can be expressed as:

$$\ln(Y') = \text{logit}\,[\pi(\underline{x})] = \ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right)$$

$$= \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p. \qquad (2)$$

In Stata, the ordinal logistic regression model is expressed in logit form as follows:

$$\ln(Y_j') = \text{logit}\,[\pi(x)] = \ln\left(\frac{\pi_j(x)}{1-\pi_j(x)}\right)$$

$$= \alpha_j + (-\beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_p X_p), \qquad (3)$$

where $\pi_j(\underline{x}) = \pi(Y \le j|x_1,x_2,\ldots x_p)$, which is the probability of being at or below category j, given a set of predictors. j = 1, 2, … J -1. α_j are the cut points, and β_1, $\beta_2 \ldots \beta_p$ are logit coefficients. This is the form of a Proportional Odds (PO) model because the odds ratio of any predictor is assumed to be constant across all categories. Similar to logistic regression, in the proportional odds model we work with the logit, or the natural log of the odds. To estimate the ln (odds) of being at or below the j^{th} category, the PO model can be rewritten as:

$$\text{logit}\,[\pi(Y \le j \mid x1,x2,\ldots x_p)] = \ln\left(\frac{\pi(Y \le j \mid x_1, x_2,\ldots x_p)}{\pi(Y > j \mid x_1, x_2,\ldots x_p)}\right)$$

$$= \alpha_j + (-\beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_p X_p) \qquad (4)$$

Thus, this model predicts cumulative logits across J -1 response categories. By transforming the cumulative logits, we can obtain the estimated cumulative odds as well as the cumulative probabilities being at or below the j^{th} category.

SAS uses a different ordinal logit model for estimating the parameters from Stata. For SAS PROC LOGISTIC (the ascending option), the ordinal logit model has the following form:

$$\text{logit}\,[\pi(Y \le j \mid x_1, x_2,\ldots x_p)] = \ln\left(\frac{\pi(Y \le j \mid x_1, x_2,\ldots x_p)}{\pi(Y > j \mid x_1, x_2,\ldots x_p)}\right)$$

$$= \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p; \qquad (5)$$

Using SAS with the descending option, the ordinal logit model can be expressed as:

$$\text{logit}\,[\pi(Y \ge j \mid x_1, x_2,\ldots x_p)] = \ln\left(\frac{\pi(Y \ge j \mid x_1, x_2,\ldots x_p)}{\pi(Y < j \mid x_1, x_2,\ldots x_p)}\right)$$

$$= \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p, \qquad (6)$$

where in both equations α_j are the intercepts, and β_1, $\beta_2 \ldots \beta_p$ are logit coefficients.

SPSS PLUM (Polytomous Universal Model) is an extension of the generalized linear model for ordinal response data. It can provide five types of link functions including logit, probit, complementary log-log, cauchit and negative log-log. Just as Stata, the ordinal logit model is also based on the latent continuous outcome variable for SPSS PLUM, it takes the same form as follows:

$$\text{logit}\,[\pi(Y \le j \mid x_1, x_2,\ldots x_p)] = \ln\left(\frac{\pi(Y \le j \mid x_1, x_2,\ldots x_p)}{\pi(Y > j \mid x_1, x_2,\ldots x_p)}\right)$$

$$= \alpha_j + (-\beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_p X_p), \qquad (7)$$

where α_j's are the thresholds, and β_1, $\beta_2 \ldots \beta_p$ are logit coefficients; j = 1, 2…J-1.

Compared to both Stata and SPSS, SAS (ascending and descending) does not negate the signs before the logit coefficients in the equations, because SAS Logistic procedure (Proc Logistic) is used to model both the dichotomous and ordinal categorical dependent variables, and the signs before the coefficients in the ordinal logit model are kept consistent with those in the binary logistic regression model.

Although the signs in the equations are positive, SAS internally changes the signs of the estimated intercepts and coefficients according to different ordering of the dependent variable (with the ascending or descending option).

Methodology

Sample

The data were collected from teachers at three middle schools and a teacher's training school in Taizhou City, Jiangsu Province, China, using a survey instrument named Teachers' Perceptions of Grading Practices (TPGP) (Liu, 2004; Liu, O'Connell & McCoach, 2006). A total of 147 teachers responded to the survey with the response rate of 73.5%. The outcome variable of interest is teachers' teaching experiences, which is an ordinal categorical variable with 1 = low, 2 = medium and 3 = high.

Explanatory variables included gender (female = 1; male = 2) and a set of scale scores from the TPGP survey instrument The instrument included five scales measuring the importance of grading, the usefulness of grading, student effort influencing grading, student ability influencing grading, and teachers' grading habits. Composite scale scores were created by taking a mean of all the items for each scale. Table 1 displays the descriptive statistics for these independent variables.

The proportional odds model was first fitted with a single explanatory variable using Stata (V. 9.2) OLOGIT. Afterwards, the full-model was fitted with all six explanatory variables. The assumption of proportional odds for both models was examined using the Brant test. Additional Stata subcommands demonstrated here included FITSTAT and LISTCEOF of Stata SPost (Long & Freese, 2006) used for the analysis of post-estimations for the models. The results of fit statistics, cut points, logit coefficients and cumulative odds of the independent variables for both models were interpreted and discussed. The same model was fit using SAS (V. 9.1.3) (ascending and descending), and SPSS (V. 13.0), and the similarities and differences of the results using all three programs were compared.

Table 1: Descriptive Statistics for All Variables, n = 147

Variable	Teaching Experience Level			
	1 n = 70 47.6%	2 n = 45 30.6%	3 n = 32 21.8%	Total n = 147 100%
% Gender (Female)	74.3%	66.7%	50%	66.7%
Importance	3.33 (.60)	3.31 (.63)	3.55 (.79)	3.37 (.66)
Usefulness	3.71 (.61)	3.38 (.82)	3.70 (.66)	3.60 (.70)
Effort	3.77 (.50)	3.79 (.46)	3.80 (.68)	3.78 (.53)
Ability	3.74 (.40)	3.75 (.54)	3.87 (.51)	3.77 (.47)
Habits	3.38 (.66)	3.57 (.66)	3.49 (.60)	3.46 (.65)

Results

Proportional Odds Model with a Single Explanatory Variable

OLOGIT is the Stata program estimating ordinal logistic regression models of ordinal outcome variable on the independent variables. In this example, the outcome variable, teaching was followed immediately by the independent variable, gender. Figure 1 displays the Stata output for the one-predictor proportional odds model.

The log likelihood ratio Chi-Square test with 1 degree of freedom, LR $\chi^2_{(1)} = 5.29$, p = .0215, indicated that the logit regression coefficient of the predictor, gender was statistically different from 0, so the full model with one predictor provided a better fit than the null model with no independent variables in predicting cumulative probability for teaching experience level. The likelihood ratio R^2_L = .0172, which is the Pseudo R^2, and is also called McFadden's R^2, suggested that the relationship between the response variable, teaching experience, and the predictor, gender was small. More measures of fit were obtained using SPost subcommand fitstat (Long & Freese, 2006). In addition to the deviance statistic and McFadden's R^2, several other types of R^2 statistics were reported (Figure2). The information measures, AIC and BIC, were used to compare either nested or non-nested models. Smaller AIC and BIC statistics indicate the better fitting model.

Figure 1: Stata Proportional Odds Model Example: Gender

```
. ologit teaching gender

Iteration 0:   log likelihood = -153.99556
Iteration 1:   log likelihood = -151.35669
Iteration 2:   log likelihood = -151.35194

Ordered logistic regression                     Number of obs   =        147
                                                LR chi2(1)      =       5.29
                                                Prob > chi2     =     0.0215
Log likelihood = -151.35194                     Pseudo R2       =     0.0172

------------------------------------------------------------------------------
    teaching |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .7587563   .3310069     2.29   0.022     .1099947    1.407518
-------------+----------------------------------------------------------------
       /cut1 |   .9043487   .4678928                     -.0127044    1.821402
       /cut2 |   2.320024   .5037074                      1.332775    3.307272
------------------------------------------------------------------------------
```

Figure 2: Measures of Fit Statistics

```
. fitstat

Measures of Fit for ologit of teaching

Log-Lik Intercept Only:       -153.996   Log-Lik Full Model:          -151.352
D(144):                        302.704   LR(1):                          5.287
                                         Prob > LR:                      0.021
McFadden's R2:                   0.017   McFadden's Adj R2:             -0.002
ML (Cox-Snell) R2:               0.035   Cragg-Uhler(Nagelkerke) R2:     0.040
McKelvey & Zavoina's R2:         0.038
Variance of y*:                  3.419   Variance of error:              3.290
Count R2:                        0.476   Adj Count R2:                   0.000
AIC:                             2.100   AIC*n:                        308.704
BIC:                          -415.918   BIC':                          -0.297
BIC used by Stata:             317.675   AIC used by Stata:            308.704
```

The estimated logit regression coefficient, β = .7588. z = 2.29, p = .022, indicating that gender had a significant effect on teacher's teaching experience level. Substituting the value of the coefficient into the formula (4), logit $[\pi(Y \leq j \mid \text{gender})] = \alpha_j + (-\beta_1 X_1)$, we calculated logit $[\pi(Y \leq j \mid \text{gender})] = \alpha_j$ - .7588 (gender). OR = $e^{(-.7588)}$ = .468, indicating that male teachers were .468 times the odds for female teachers of being at or below at any category, i.e., female teachers were more likely than male teachers to be at or below a particular category, because males were coded as 2 and girls as 1.

The results table reports two cut-points: _cut1 and_cut2. These are the estimated cut-points on the latent variable, Y*, used to differentiate the adjacent levels of categories of teaching experiences. When the response category is 1, the latent variable falls at or below the first cut point, α_1. When the response category is 2, the latent variable falls between the first cut point α_1 and the second cut point α_2, and when the response category reaches 3 if the latent variable is at or beyond the second cut point α_2.

To estimate the cumulative odds being at or below a certain category, j for gender, the logit form of proportional odds model was used, logit $[\pi(Y \leq j \mid \text{gender})] = \alpha_j$ - .7588 (gender). For example, when $Y \leq 1$, α_1, .9043 is the first cut point for the model. Substituting it into the formula (4) results in logit $[\pi(Y \leq j \mid \text{gender})]$ = .9043 - .7588 (gender). For girls (x = 1), logit $[\pi(Y \leq 1 \mid \text{gender})]$ = .1455. By exponentiating the logit, the odds for female teachers of being at or below experience category 1 is calculated, $e^{.1455}$= 1.157. For male teachers (x = 2), logit $[\pi(Y \leq 1 \mid \text{gender})]$ = .9043 - .7588*2 = -.6133, so the odds for male teachers being at or below teaching experience category 1, $e^{-.6133}$=.542. Odds ratio of male teachers versus female teachers = .542/1.157 = .468. Transforming the cumulative odds, results in the cumulative probabilities by using p = odds/(1+odds).

The Stata program brant was used to test the proportional odds assumption. Brant (1990) proposed a test of proportional odds assumption for the ordinal logistic model by examining the separate fits to the underlying binary logistic models. A non-significant omnibus test indicates that the proportional odds assumption is not violated. It also provides tests for each individual independent variable. When only one independent variable exists in the model, the results of the omnibus test and individual test are the same. The Brant test of parallel regression assumption yields χ^2_1 = .40 (p > .527), indicating that the proportional odds assumptions for the full-model was upheld. This suggests that the effect of gender, the explanatory variable, was constant across separate binary models fit to the cumulative cut points. Figure 3 also shows the estimated coefficient from j-1 binary logistic regression models. Each logistic regression model estimates the probability of being at or beyond teaching experience level *j*.

Figure 3: Brant Test of Parallel Regression (Proportional Odds) Assumption

```
. brant, detail
Estimated coefficients from j-1 binary regressions

                y>1         y>2
gender    .66621777   .91021169
 _cons   -.78882009  -2.5443422

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+---------------------------
         All |      0.40    0.527     1
-------------+---------------------------
      gender |      0.40    0.527     1
-----------------------------------------

A significant test statistic provides evidence that the parallel regression
assumption has been violated.
```

The Proportional Odds Model can also estimate the ln(odds) of being at or beyond category j, given a set of predictors. Again, these ln(odds) can be transformed into the cumulative odds, and cumulative probabilities as well. For example, the cumulative probability of a teacher's teaching experience can be estimated at or beyond category 3, $P(Y \geq 3)$, which is the complementary probability when $Y \leq 2$, at or beyond category 2, $P(Y \geq 2)$, and $P(Y \geq 1)$, which equals 1.

In Stata, when estimating the odds of being beyond category j, or at or beyond j+1, the sign of the cut points needs to be reversed and their magnitude remain unchanged because the cut points were estimated from the right to the left of the latent variable, Y*, that is, from the direction when Y = 3 approaches Y = 1. Therefore, two cut points from right to left turn to -2.32 and - .904. When the predictor is dichotomous, a positive sign of the logit coefficient indicates that it is more likely for the group (x = 1) to be at or beyond a particular category than for the relative group (x = 0). When the predictor is continuous, a positive coefficient indicates that when the value of the predictor variable increases, the probability of being at or beyond a particular category increases.

Using Stata syntax listcoef, the odds of being at or beyond a particular category at 2.136 can be obtained, which was constant across all cumulative categories. It also indicated that male teachers were 2.136 times the odds for female teachers of being at or beyond any category, i.e., male teachers were more likely than female teachers to be at or beyond a particular category. Figure 4 displays the results of Stata listcoef. Adding option percent after listcoef, the result of percentage change in odds of being at or beyond a particular category can be obtained when the predictor, gender, goes from males (x = 2) to females (x = 1).

Proportional Odds Model with Six Explanatory Variables

Next, a proportional odds model was fit with eight explanatory variables, which is referred to as the Full Model. Figure 5 displays the results for the fitting of the full model with six explanatory variables.

Before interpreting the results of the full model, the assumption of proportional odds was first examined. The Stata brant command provides the results of the Brant test of parallel regression (Proportional Odds) assumption for the full model with six predictors and tests for each independent variable. It also provides the estimated coefficient from j-1 binary logistic regression models results of two separate binary logistic regression models. The data are dichotomized according to the cumulative probability pattern so that each logistic

Figure 4: Results of Stata listcoef

```
. listcoef, help

ologit (N=147): Factor Change in Odds

  Odds of: >m vs <=m

------------------------------------------------------------------------------
    teaching |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------------
      gender |   0.75876    2.292   0.022   2.1356   1.4318     0.4730
------------------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

regression model estimates the probability of being at or beyond teaching experience level *j*. For the omnibus Brant test, $\chi^2_6 = 8.10$, p = .230, indicating that the proportional odds assumptions for the full-model was upheld. Examining the Brant tests for each individual independent variable indicated that the Brant test of the assumption of parallel regression (proportional odds) were upheld for gender, importance, effort, ability and habits. For usefulness, the Brant test, $\chi^2_1 = 4.03$, p = .045, which is very close to .05, therefore, it may also be concluded that the PO assumption for this variable is nearly upheld. Checking the estimated coefficients for each independent variable across two binary logistic regression models shows that the logit coefficients for all the variables were similar across two binary logistic models, supporting the results of the Brant test of proportional odds assumption.

The log likelihood ratio Chi-Square test, LR $\chi^2_{(6)} = 13.738$, p = .033, indicating that the full model with six predictor provided a better fit than the null model with no independent variables in predicting cumulative probability for teaching experience. The likelihood ratio R^2_L = .045, much larger than that of the gender-only model, but still small, suggesting that the relationship between the response variable, teaching experience, and six predictors, was still small. Compared with the gender-only model, all R^2statistics of the full-model shows improvement (see Figure 7).

Figure 6: Brant Test of Parallel Regression (Proportional Odds) Assumption

```
. brant, detail

Estimated coefficients from j-1 binary regressions

                    y>1         y>2
    gender    .74115294   .86086025
importance    .64416122   .46874536
usefulness   -.94566294  -.19259753
    effort    .09533898  -.03621639
   ability    .26862373   .68349765
    habits    .48959286  -.02795948
     _cons   -2.7097459  -5.7522624

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+---------------------------
         All |      8.10    0.231     6
-------------+---------------------------
      gender |      0.08    0.772     1
  importance |      0.24    0.622     1
  usefulness |      4.03    0.045     1
      effort |      0.10    0.746     1
     ability |      0.66    0.418     1
      habits |      2.15    0.142     1
----------------------------------------

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

Figure 7: Measure of Fit Statistics for Full-Model

```
. fitstat

Measures of Fit for ologit of teaching

Log-Lik Intercept Only:       -153.996   Log-Lik Full Model:          -147.127
D(139):                        294.253   LR(6):                         13.738
                                         Prob > LR:                      0.033
McFadden's R2:                   0.045   McFadden's Adj R2:             -0.007
ML (Cox-Snell) R2:               0.089   Cragg-Uhler(Nagelkerke) R2:     0.102
McKelvey & Zavoina's R2:         0.098
Variance of y*:                  3.646   Variance of error:              3.290
Count R2:                        0.429   Adj Count R2:                  -0.091
AIC:                             2.111   AIC*n:                        310.253
BIC:                          -399.417   BIC':                          16.205
BIC used by Stata:             334.177   AIC used by Stata:            310.253
```

The Stata listcoef command (Figure 8) produced more detailed results of logit coefficients and cumulative odds (exponentiated coefficients). For the proportional odds model, interpretation of cumulative odds is independent on the ancillary parameters (cut points) because they are constant across all levels of the response variable.

The effects of the independent variables can be interpreted in several ways, including how they contribute to the odds and their probabilities of being at or beyond a particular category. They can also be interpreted as how these variables contribute to the odds of being at or below a particular category, if the sign is reversed before the estimated logit coefficients and corresponding cumulative odds are computed. In terms of odds ratios, male teachers were 2.241 times the odds for female teachers to be at or beyond a particular category (OR=2.241), after controlling the effects of other predictors in the model. The usefulness of grading with a corresponding OR significantly less than 1.0 has significant negative effects in the model. These cumulative odds are associated with a teacher being in lower teaching experience categories rather than in higher categories. For a one unit increase in the usefulness of grading, the odds ratio of being in higher teaching experience categories versus lower categories was .53 times lower, after controlling for the effects of other variables. However, variables whose corresponding ORs are significantly greater than 1.0 have significant positive effects on the response variable in the model. For example, the importance of grading (OR=1.778) had a positive effect on teachers being in higher teaching experience categories. For a one unit increase in the importance of grading, the odds ratio of being in higher teaching experience categories versus lower categories was 1.778 times greater, given the effects of other predictors are held constant. Variables such as student ability and teacher's grading habits, whose corresponding ORs were greater than 1.0, but were not statistically significant, had positive effects on the response variable, but these effects may be due to chance and need further investigation. Independent variables with ORs close to 1.0 have no effect on the response variable. For example, student effort influencing grading was not associated with teaching experience in this model (OR=1.0266, p=.946).

Comparison of Results of a Single-Variable PO Model Using Stata, SAS, and SPSS

Table 2 shows a comparison of the results for Stata OLOGIT with results from SAS PROC LOGISTIC with the ascending and descending options, and SPSS PLUM. The similarities and differences between these results should be noted, otherwise, it could be misleading to interpret the results in the same way, disregarding their different parameterizations. In estimating proportional odds models, Stata sets the intercept to 0, and estimates the cut points, while SAS ascending

Figure 8: Results of Logit Coefficient, Cumulative Odds, and Percentage Change in Odds

```
. listcoef, help

ologit (N=147): Factor Change in Odds

  Odds of: >m vs <=m

------------------------------------------------------------------------------
    teaching |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------------
      gender |   0.80695    2.318   0.020   2.2411   1.4648     0.4730
  importance |   0.57547    1.897   0.058   1.7780   1.4601     0.6578
  usefulness |  -0.63454   -2.322   0.020   0.5302   0.6402     0.7029
      effort |   0.02625    0.068   0.946   1.0266   1.0140     0.5283
     ability |   0.34300    0.825   0.409   1.4092   1.1752     0.4707
      habits |   0.31787    1.088   0.277   1.3742   1.2282     0.6466
------------------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

estimates the intercepts and set the cut points to 0. Comparing Stata with SAS (ascending), the different choice of parameterization does not influence the magnitude of cut points (or intercepts) and coefficients. However, it does determine the sign before these estimates.

When estimating the odds of being at or below a response category, the estimates for the cut points using Stata are the same as the intercepts using SAS ascending in both sign and magnitude. The first cut point, α_1 in Stata estimation is the same as the first intercept α_1 in SAS ascending estimation, because there is no first intercept α_1 in Stata estimation. Using Stata and SAS (the ascending option), the estimated logit coefficients are the same in magnitude but are opposite in sign. Using Stata, the estimated logit coefficient $\beta = .759$. Substituting it into the logit form (4), we get logit $[\pi(Y \leq j \mid \text{gender})] = \alpha_j - (.759)*(\text{gender}) = \alpha_j - .759*(\text{gender})$. OR = $e^{(-.759)} = .468$, indicating that male teachers were .468 times the odds for female teachers of being at or below at any category, that is, female teachers were more likely than male teachers to be at or below a particular teaching experience level. Using SAS ascending, the estimated logit coefficient, $\beta = -.759$. Substituting it into its corresponding logit form (5) results in the same equation: logit $[\pi(Y \leq j \mid \text{gender})] = \alpha_j - .759*(\text{gender})$. Therefore, the same results of estimated cumulative odds and cumulative probability were obtained using Stata and SAS ascending.

Comparing the results of the proportional odds model using Stata and SAS with the descending option, it was found that estimated cut points for Stata and the estimated intercepts for SAS descending are the same in magnitude but are opposite in sign. Using Stat and SAS descending, the estimated logit coefficients are the same in both magnitude and sign. To estimate the odds of being at or beyond a particular teaching experience level using Stata, it is only necessary to reverse the sign of the estimated cut points. The estimated logit coefficient is $\beta = .759$. Exponentiating this results in $e^{(.759)} = 2.136$, indicating male teachers are 2.136 times greater than female teachers to be at or beyond a particular category. In other words, female teachers are less likely than male teachers to be at or beyond a certain category.

Table 2: Results of Proportional Odds Model with a Single Variable Using Stata, SAS (Ascending and descending) and SPSS: A Comparison

Model Estimates	STATA	SAS (Ascending)	SAS (Descending)	SPSS
	P(Y ≤ j)	P(Y ≤ j)	P(Y ≥ j)	P(Y ≤ j)
Cutpoints (Stata)/ Intercept (SAS)/ Threshold (SPSS)	_cut1(α_1) = .904	α_1 = .904	α_3 = -2.32	α_1 = -.613
	_cut2(α_2) = 2.32	α_2 = 2.32	α_2 = -.904	α_2 = .803
Gender (Male = 2)	.759	-.759	.759	0
Gender (Female = 1)				-.759
LR R^2	.017	.017	.017	.017
Brant Test (Omnibus Test)[a]	χ^2_1 = .40 (p > .527)			
Score Test[b]		χ^2_1 = .4026 (p = .5258)	χ^2_1 = .4026 (p = .5258)	χ^2_1 = .392 (p > .530)
Model Fit	LR $\chi^2_{(1)}$ = 5.29, p = .0215	LR $\chi^2_{(1)}$ = 5.29, p = .0215	LR $\chi^2_{(1)}$ = 5.29, p = .0215	LR $\chi^2_{(1)}$ = 5.287, p = .021

a. Brant test for proportional odds assumption.
b. Score test for proportional odds assumption.

Using Stata and SPSS, when estimating the effects of predictors on the log odds of being at or below a certain category of the outcome variable, the sign before the coefficients are both minus rather than plus. In other words, the effects of predictors are subtracted from the cut points or thresholds. SPSS PLUM labels the estimated logits for the predicator variables LOCATION. When the predicator variable is continuous, the estimated logit coefficients are the same as those estimated by Stata OLOGIT in both magnitude and sign. However, SPSS PLUM is different from Stata OLOGIT in this aspect: when the predictor variable is categorical, for example gender, with 1 = female and 2 = male, the estimated coefficient is only displayed for the category with smaller value, i.e., when gender = 1. The category with larger value, gender = 2, is the reference category, and has an estimate of 0. If gender is coded with 1 = female and 0 = male, the estimated coefficient is displayed for the case when gender = 0, and the estimated coefficient for female (gender = 1) is 0. Using SPSS PLUM, the estimated logit coefficient, β = -.759 for the case when female = 1, and β = 0 for the case when male = 2. Substituting it into the logit form (7) results in logit [π(Y ≤ j | gender)] = α_j –(–.759)*(gender) = α_j +.759*(gender). By exponentiating, OR = $e^{(.759)}$ = 2.136, indicating that female teachers are

2.136 times the odds for male teachers of being at or below at a particular teaching experience level. This result is equivalent to that of Stata.

The parameter estimation for the categorical predictor in SPSS PLUM makes the threshold values in the ordinal logit model different from those estimated by Stata OLOGIT. These differences can be observed in the results of the proportional odds model using Stata, SAS (ascending and descending), and SPSS (Table 2). In SPSS PLUM, the threshold estimates are for the case when gender = 2 (male teachers), while in Stata and SAS, the cut points or intercepts are for case when gender = 1 (female students).

Equivalent results of estimated logit can be obtained using different estimates of cutpoints (thresholds) and logit coefficients fitted by Stata and SPSS. For example, using SPSS, the predicted logit for male teachers (gender = 2) of being at or below teaching experience level 1, logit $[\pi(Y \leq 1 \mid \text{gender})] = \alpha_1 - 0*(\text{gender}) = -.613 + 0*(2) = -.613$; the predicted logit for female teachers (gender = 1) of being at or below teaching experience level 1, logit $[\pi(Y \leq 1 \mid \text{gender})] = \alpha_1 - (-.759)*(\text{gender}) = -.613 + .759*1 = .146$. Using Stata, the predicted logit for male teachers (gender = 2) of being at or below teaching experience level 1, logit $[\pi(Y \leq 1 \mid \text{gender})] = \alpha_1 - (.759)*(\text{gender}) = .904 - .759*2 = -.614$; the predicted logit for female teachers (gender =1) of being at or below teaching experience level 1, logit $[\pi(Y \leq 1 \mid \text{gender})] = \alpha_1 - (.759)*(\text{gender}) = .904 - .759*1 = .145$.

To test the proportional odds assumption, Stata uses the Brant test of parallel regression assumption with the result $\chi^2_1 = .40$ p > .527; SAS uses ascending and descending score test and has the same results $\chi^2_1 = .4026$, p = .5258; SPSS uses a test of parallel lines with the result $\chi^2_1 = .392$, p > .530. All tests produce similar results in that the proportional odds model assumption is upheld. Across the models, the omnibus likelihood ratio tests produce the same results, indicating the proportional odds model with one variable (gender) has better fit than the null model. Features of the ordinal logistic regression analysis using Stata, SAS and SPSS are shown and compared in Table 3.

Conclusion

In this article, the use of proportional odds models was illustrated to predict teachers' teaching experience level from a set of measures of teachers' perceptions of grading practices. A single independent variable model and a full-model with six independent variables were fitted and compared. The assumptions of proportional odds for both models were examined. It was found that the assumption of proportional odds for both the single-variable model and the full-model was upheld.

Results from the proportional odds model revealed that the usefulness of grading had a negative effect on the prediction of teaching experience level (OR = .53), while the importance of grading practices had a positive effect on the experience level (OR = 1.78), after controlling for the effects of other variables. Although student effort influencing teachers' grading practices (OR = 1.41) and teachers' grading habits (OR = 1.37) had positive effects on teaching experience level, these effects were not found to be significant. Compared to male teachers, female teachers were more likely to be at or below a particular category, or in other words, males were more likely to be at or beyond an experience level. Student effort influencing grading was not associated with teachers' teaching experience level in the model.

These findings suggest that teachers with longer teaching experience tended to feel the grading practices are more important than the teachers with fewer years of teaching. However, teachers with longer teaching experiences tended to doubt the usefulness of grading in their teaching; this may be due in part to their requirement of conducting test-oriented teaching in China. In addition, the gender difference suggests that female teachers were more easily categorized as inexperienced teachers; this may be due to greater numbers of female students receiving the opportunities of higher education in recent years and their choosing teaching as their profession. The frequencies of new female teachers are currently greater than those of new male teachers in China.

Comparing the results using Stata and SAS, it was found that both packages produced the same or similar results in model fit statistics,

Table 3: Feature Comparisons of the Ordinal Logistic Regression Analysis Using Stata, SAS and SPSS

	STATA	SAS	SPSS
Model Specification			
Cutpoints/Thresholds	✓		✓
Intercept		✓	
Test Hypotheses of Logit Coefficients	✓	✓	✓
Maximum Likelihood Estimates			
Odds Ratio	✓	✓	
z-statistic or Wald Test for Parameter Estimate	✓		✓
Chi-square Statistic for Parameter Estimate		✓	
Confidence Interval for Parameter Estimate	✓		✓
Fit Statistics			
Loglikelihood	✓	✓	✓
Goodness-of-Fit Test	✓	✓	✓
Pseudo R-Square	✓	✓	✓
Test of PO Assumption			
Omnibus Test of Assumption of Proportional Odds	✓	✓	✓
Test of Assumption of Proportional Odds for Individual Variables	✓		
Association of Predicted Probabilities and Observed Responses		✓	

and the test of proportional odds assumption. The estimated coefficients and cut points (thresholds) were the same in magnitude but may be reversed in sign. Comparing the results using Stata and SPSS, it was found that although the ordinal logit models are based on latent continuous response variables for both packages, SPSS PLUM estimated the logit coefficient for the category with smaller value when the predictor variable was categorical, and thus the estimated thresholds were different from those estimated by Stata. Researchers should understand the differences of parameterization of ordinal logistic models using Stata and other statistical packages. Researchers should pay attention to the sign before the estimated logit coefficients and the cut points in the model, and exercise caution in interpreting the results.

In educational research, ordinal categorical data is frequently used and researchers need to understand and be familiar with the ordinal logistic regression models dealing with the internally ordinal outcome variables. In some situations, Ordinary Least Squares (OLS) techniques may be used for preliminary analysis of such data by treating the ordinal scale variable as continuous. However, ignoring the discrete ordinal nature of the variable would cause the analysis lose some useful information and could lead to misleading results. Therefore, it is crucial for researchers to use the most appropriate models to analyze ordinal categorical dependent variables. In addition, the role of any statistical software

package is a tool for researchers. The choice of software is the preference of researchers; it is therefore not the purpose of the study to suggest which one is the best for ordinal logistic regression analysis. This demonstration clarifies some of the issues that researchers must consider in using different statistical packages when analyzing ordinal data.

References

Agresti, A. (2002). *Categorical data analysis* (*2nd Ed.*). NY: John Wiley & Sons.

Agresti, A. (1996). *An introduction to categorical data analysis*. NY: John Wiley & Sons.

Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute, Inc.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*, 1323-1333.

Armstrong, B. B., & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, *129*(*1*), 191-204.

Bender, R., & Benner, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, *42*(*6*), 677-699.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*, 1171-1178.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

Greene, W. H. (2003). *Econometric analysis* (*5th Ed.*). Upper Saddle River, NJ: Prentice Hall.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (*2nd Ed.*). NY: John Wiley & Sons.

Liu, X. (2004). *A validation of teachers' perceptions of grading practices*. Paper presented at the October Annual Conference of the Northeastern Educational Research Association (NERA), Kerhonkson, NY.

Liu, X., O'Connell, A. A., & McCoach D. B. (2006). *The initial validation of teachers' perceptions of grading practices.* Paper presented at the April 2006 Annual Conference of the American Educational Research Association (AERA), San Francisco, CA.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (*2nd Ed.*). Texas: Stata Press.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, *B*, *42*, 109-142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (*2nd Ed.*). London: Chapman and Hall.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

O'Connell, A. A., (2000). Methods for modeling ordinal outcome variables. *Measurement and Evaluation in Counseling and Development*, *33*(*3*), 170-193.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: SAGE.

O'Connell, A. A., Liu, X., Zhao, J., & Goldstein, J. (2006). *Model Diagnostics for proportional and partial proportional odds models*. Paper presented at the April 2006 Annual Conference of the American Educational Research Association (AERA), San Francisco, CA.

Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego, CA: Academic Press.

Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 646-658

1538 – 9472/09/$95.00

JMASM ALGORITHMS & CODE

JMASM28: Gibbs Sampling for 2PNO Multi-unidimensional Item Response Theory Models (Fortran)

Yanyan Sheng Todd C. Headrick
Southern Illinois University Carbondale

A Fortran 77 subroutine is provided for implementing the Gibbs sampling procedure to a multi-unidimensional IRT model for binary item response data with the choice of uniform and normal prior distributions for item parameters. In addition to posterior estimates of the model parameters and their Monte Carlo standard errors, the algorithm also estimates the correlations between distinct latent traits. The subroutine requires the user to have access to the IMSL library. The source code is available at http://www.siuc.edu/~epse1/sheng/Fortran/MUIRT/GSMU2.FOR. An executable file is also provided for download at http://www.siuc.edu/~epse1/sheng/Fortran/MUIRT/EXAMPLE.zip to demonstrate the implementation of the algorithm on simulated data.

Key words: multi-unidimensional IRT model, two-parameter normal ogive model, MCMC, Gibbs sampling, Fortran.

Introduction

Modeling the interaction of a person's trait and the test at the item level for binary response data, the conventional item response theory (IRT) models rely on a strong assumption of unidimensionality. That is, each test item is designed to measure some facet of a unified latent trait. However, psychological processes have consistently been found to be more complex and an increasing number of educational measurements assess a person on more than one latent trait. In the situations when a test consists of several subtests with each measuring one latent trait, the multi-unidimensional IRT models have been found to be more appropriate than the unidimensional models (Sheng & Wikle, 2007), as they allow inferences to be made about a person for each distinct trait being measured.

For the two-parameter normal ogive (2PNO) multi-unidimensional model, the probability of person i obtaining a correct response for item j in subtest v, where $i = 1,...,n$, $j = 1,...,k_v$, $v = 1,...,m$, and $K = \sum_v k_v$, is defined as

$$P(y_{vij} = 1) = \Phi(\alpha_{vj}\theta_{vi} - \gamma_{vj}) = \int_{-\infty}^{\alpha_{vj}\theta_{vi} - \gamma_{vj}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \quad (1)$$

(e.g., Lee, 1995; Sheng & Wikle, 2007), where α_{vj} and θ_{vi} are scalar parameters representing the item discrimination and the continuous person trait in the vth latent dimension, and γ_{vj} is a scalar parameter indicating the location in that dimension where the item provides maximum information. To estimate both item and person parameters simultaneously, Markov

Yanyan Sheng is an Associate Professor of Measurement and Statistics. Her areas of research interest are psychometrics, IRT and Bayesian hierarchical models, adaptive testing. Email: ysheng@siu.edu. Todd C. Headrick is Professor of Statistics. Address: Section on Statistics and Measurement, Department of EPSE, 222-J Wham Building, Mail Code 4618, Southern Illinois University-Carbondale, IL, 62901. His areas of research interest are statistical computing, nonparametric statistics, and optimization. Email: headrick@siu.edu.

chain Monte Carlo (MCMC; e.g., Chib & Greenberg, 1995) techniques are used to summarize the posterior distributions that arise in the context of the Bayesian prior-posterior framework (Carlin & Louis, 2000; Chib & Greenberg, 1995; Gelfand & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 2003; Tanner & Wong, 1987). Lee (1995) applied Gibbs sampling (Casella & George, 1992; Gelfand & Smith, 1990; Geman & Geman, 1984), an MCMC algorithm, to the 2PNO multi-unidimensional model and illustrated the model parameterization by adopting non-informative priors for item parameters.

Due to the reasons that informative priors are desirable in some applications in the Bayesian framework, and MCMC is computational demanding (see Sheng & Headrick, 2007, for a description of the problems), this study focuses on using Fortran, the fastest programming language for numerical computing (Brainerd, 2003) to implement the procedure. In particular, the paper provides a Fortran subroutine that obtains the posterior estimates and Monte Carlo standard errors of estimates for the item and person parameters in the 2PNO multi-unidimensional IRT model, as well as the posterior estimates of the correlations between the distinct latent traits. The subroutine allows the user to specify non-informative and informative priors for item parameters.

Methodology

The Gibbs Sampling Procedure

To implement Gibbs sampling to the 2PNO multi-unidimensional IRT model defined in (1), a latent continuous random variable Z is introduced so that $Z_{vij} \sim N(\alpha_{vj}\theta_{vi} - \gamma_{vj}, 1)$ (Albert, 1992; Lee, 1995; Tanner & Wong, 1987). Next, denote each person's latent traits for all items as $\boldsymbol{\theta}_i = (\theta_{1i}, ..., \theta_{mi})'$, which is assumed to have a multivariate normal (MVN) distribution, $\boldsymbol{\theta}_i \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a correlation matrix, and ρ_{st} is the correlation between θ_{si} and θ_{ti}, $s \neq t$, on the off diagonals. It may be noted that the unidimensional IRT model is a special case of the multi-unidimensional model where $\rho_{st} = 1$ for all s, t. Then, an unconstrained covariance matrix $\boldsymbol{\Sigma}^*$ is introduced (Lee, 1995), where $\boldsymbol{\Sigma}^* = [\sigma_{ij}]_{m \times m}$, so that the correlation matrix $\boldsymbol{\Sigma}$ can be easily transformed from $\boldsymbol{\Sigma}^*$ using $\rho_{st} = \frac{\sigma_{st}}{\sigma_s \sigma_t} (s \neq t)$. A non-informative prior is assumed for $\boldsymbol{\Sigma}^*$ so that $p(\boldsymbol{\Sigma}^*) \propto |\boldsymbol{\Sigma}^*|^{-\frac{m+1}{2}}$. Hence, with prior distributions assumed for $\boldsymbol{\xi}_{vj}$, where $\boldsymbol{\xi}_{vj} = (\alpha_{vj}, \gamma_{vj})'$, the joint posterior distribution for $(\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Z}, \boldsymbol{\Sigma}^*)$ is

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Z}, \boldsymbol{\Sigma}^* | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{Z}) p(\mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\xi}) p(\boldsymbol{\theta} | \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}^*). \quad (2)$$

where $f(\mathbf{y} | \mathbf{Z})$ is the likelihood function.

With non-informative priors for α_{vj} and γ_{vj} so that $\alpha_{vj} > 0$ and $p(\gamma_{vj}) \propto 1$, the full conditional distributions of Z_{vij}, $\boldsymbol{\theta}_i$, $\boldsymbol{\xi}_{vj}$ and $\boldsymbol{\Sigma}^*$ can be derived in closed forms as follows:

$$Z_{vij} | \bullet \sim \begin{cases} N_{(0,\infty)}(\alpha_{vj}\theta_{vi} - \gamma_{vj}, 1), \text{ if } y_{vij} = 1 \\ N_{(-\infty,0)}(\alpha_{vj}\theta_{vi} - \gamma_{vj}, 1), \text{ if } y_{vij} = 0 \end{cases}; \quad (3)$$

$$\boldsymbol{\theta}_i | \bullet \sim N_m((\mathbf{A}'\mathbf{A} + \boldsymbol{\Sigma})^{-1}\mathbf{A}'\mathbf{B}, (\mathbf{A}'\mathbf{A} + \boldsymbol{\Sigma})^{-1}), \quad (4)$$

where $\mathbf{A}_{(K \times m)} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\alpha}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\alpha}_m \end{pmatrix}$ and

$$\mathbf{B}_{(K \times 1)} = \begin{pmatrix} \mathbf{Z}_{1i} + \boldsymbol{\gamma}_1 \\ \mathbf{Z}_{2i} + \boldsymbol{\gamma}_2 \\ \vdots \\ \mathbf{Z}_{mi} + \boldsymbol{\gamma}_m \end{pmatrix}, \ \boldsymbol{\alpha}_v = (\alpha_{v1}, ..., \alpha_{v,k_v})',$$

$$\mathbf{Z}_{vi} = (Z_{vi1}, ..., Z_{nik_v})', \ \boldsymbol{\gamma}_v = (\gamma_{v1}, ..., \gamma_{v,k_v})';$$

$$\boldsymbol{\xi}_{vj} | \bullet \sim N_2((\mathbf{x}_v'\mathbf{x}_v)^{-1}\mathbf{x}_v'\mathbf{Z}_{vj}, (\mathbf{x}_v'\mathbf{x}_v)^{-1}) I(\alpha_{vj} > 0), \quad (5)$$

where $\mathbf{x}_v = [\theta_v, -1]$;

$$\boldsymbol{\Sigma}^* \mid \bullet \sim W^{-1}(\mathbf{S}^{-1}, n) \qquad (6)$$

(an inverse Wishart distribution), where $\mathbf{S} = \sum_{i=1}^{n} (c\boldsymbol{\theta}_i)(c\boldsymbol{\theta}_i)'$ and

$$c = \begin{pmatrix} (\prod_j \alpha_{1j})^{\frac{1}{k_1}} & 0 & \cdots & 0 \\ 0 & (\prod_j \alpha_{2j})^{\frac{1}{k_2}} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & (\prod_j \alpha_{mj})^{\frac{1}{k_m}} \end{pmatrix}$$

(see Lee, 1995 for a detailed derivation).

Alternatively, conjugate priors can be assumed for α_{vj} and γ_{vj} so that $\alpha_{vj} \sim N_{(0,\infty)}(\mu_{\alpha_v}, \sigma^2_{\alpha_v})$, $\gamma_{vj} \sim N(\mu_{\gamma_v}, \sigma^2_{\gamma_v})$. In this case, the full conditional distribution of $\boldsymbol{\xi}_{vj}$ is derived to be

$$\boldsymbol{\xi}_{vj} \mid \bullet \sim N_2((\mathbf{x}_v'\mathbf{x}_v + \boldsymbol{\Sigma}^{-1}_{\xi_v})^{-1}(\mathbf{x}_v'\mathbf{Z}_{vj} + \boldsymbol{\Sigma}^{-1}_{\xi_v}\boldsymbol{\mu}_{\xi_v}), (\mathbf{x}'\mathbf{x} + \boldsymbol{\Sigma}^{-1}_{\xi_v})^{-1}) I(\alpha_{vj} > 0) \qquad (7)$$

where $\boldsymbol{\mu}_{\xi_v} = (\mu_{\alpha_v}, \mu_{\gamma_v})'$ and $\boldsymbol{\Sigma}_{\xi_v} = \begin{pmatrix} \sigma^2_{\alpha_v} & 0 \\ 0 & \sigma^2_{\gamma_v} \end{pmatrix}$.

Hence, with starting values $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\xi}^{(0)}$, and $\boldsymbol{\Sigma}^{(0)}$, observations ($\mathbf{Z}^{(l)}$, $\boldsymbol{\theta}^{(l)}$, $\boldsymbol{\xi}^{(l)}$, $\boldsymbol{\Sigma}^{(l)}$) can be simulated using the Gibbs sampler by iteratively drawing from their respective full conditional distributions specified in (3), (4), (5), and (6) (or (3), (4), (7), and (6)). In particular, to go from ($\mathbf{Z}^{(l-1)}$, $\boldsymbol{\theta}^{(l-1)}$, $\boldsymbol{\xi}^{(l-1)}$, $\boldsymbol{\Sigma}^{(l-1)}$) to ($\mathbf{Z}^{(l)}$, $\boldsymbol{\theta}^{(l)}$, $\boldsymbol{\xi}^{(l)}$, $\boldsymbol{\Sigma}^{(l)}$), it takes four transition steps:

1. Draw $\mathbf{Z}^{(l)} \sim p(\mathbf{Z} \mid \mathbf{y}, \boldsymbol{\theta}^{(l-1)}, \boldsymbol{\xi}^{(l-1)})$;
2. Draw $\boldsymbol{\theta}^{(l)} \sim p(\boldsymbol{\theta} \mid \mathbf{Z}^{(l)}, \boldsymbol{\xi}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)})$;
3. Draw $\boldsymbol{\xi}^{(l)} \sim p(\boldsymbol{\xi} \mid \mathbf{Z}^{(l)}, \boldsymbol{\theta}^{(l)})$;
4. Draw $\boldsymbol{\Sigma}^{*(l)} \sim p(\boldsymbol{\Sigma}^* \mid \boldsymbol{\theta}^{(l)}, \boldsymbol{\xi}^{(l)})$, and transform $\boldsymbol{\Sigma}^{*(l)}$ to $\boldsymbol{\Sigma}^{(l)}$.

This iterative procedure produces a sequence of samples for the model parameters ($\boldsymbol{\theta}^{(l)}, \boldsymbol{\xi}^{(l)}$) and the hyperparameter $\boldsymbol{\Sigma}^{(l)}$, l = 0, …, L. To reduce the effect of the starting values, early iterations in the Markov chain are set as burn-ins to be discarded. Samples from the remaining iterations are then used to summarize the posterior density of item parameters $\boldsymbol{\xi}$, distinct person trait parameters $\boldsymbol{\theta}$, and the correlation matrix $\boldsymbol{\Sigma}$. As with standard Monte Carlo, the posterior means of all the samples collected after burn-in are considered as estimates of the true parameters $\boldsymbol{\xi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\Sigma}$.

However, the Monte Carlo standard errors cannot be calculated using the sample standard deviations because subsequent samples in each Markov chain are autocorrelated (e.g., Patz & Junker, 1999). One approach to calculating them is through batching (Ripley, 1987). That is, with a long chain of samples being separated into contiguous batches of equal length, the Monte Carlo standard error for each parameter is then estimated to be the standard deviation of these batch means. The Monte Carlo standard error of estimate is hence a ratio of the Monte Carlo standard error and the square root of the number of batches.

The Fortran Subroutine

The subroutine initially sets the starting values for the model parameters, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, and the hyperparameter $\boldsymbol{\Sigma}$, so that $\theta^{(0)}_{vi} = 0$, $\alpha^{(0)}_{vi} = 2$, $\gamma^{(0)}_{vi} = -\Phi^{-1}(\sum_i y_{vij}/n)\sqrt{5}$ (Albert, 1992), and $\boldsymbol{\Sigma}^{(0)} = \mathbf{I}$, with $\mathbf{I}$ being the identity matrix. It then iteratively draws random samples for $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}^*$ from their respective full conditional distributions specified in (3), (4) and (6). Samples for $\boldsymbol{\xi}_{vj}$ are simulated either from (5), where uniform priors are assumed for $\boldsymbol{\xi}_{vj}$, or from (7), where normal priors are adopted with

$\mu_{\alpha_v} = \mu_{\gamma_v} = 0$ and $\sigma^2_{\alpha_v} = \sigma^2_{\gamma_v} = 1$. The algorithm continues until all the *L* samples are simulated. It then discards the early burn-in samples, and computes the posterior estimates and Monte Carlo standard errors of estimates for the model parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, as well as the hyperparameter $\boldsymbol{\Sigma}$, using batching.

For example, consider binary responses of 2,000 persons to a total of 16 test items, which are further divided into two subtests so that the first half measures one latent trait and the second half measures another (i.e., n = 2,000, m = 2, k_1 = 8, k_2 = 8, and K = 16). Three dichotomous (0-1) data matrices were simulated from the item parameters shown in the first column of Tables 1 and 2, so the actual correlation (ρ) between the two distinct latent traits (θ_1, θ_2) was set to be 0.2, 0.5 and 0.8, respectively. The Gibbs sampler was implemented to each data set so that 10,000 samples were simulated with the first 5,000 taken to be burn-in. The remaining 5,000 samples were separated into 5 batches, each with 1,000 samples.

With the uniform or the normal prior distributions described previously, two sets of the posterior means for α_v, γ_v, and ρ as well as their Monte Carlo standard errors were obtained for each simulated data and are displayed in the rest of the tables. Note that in all the three simulated situations, item parameters were estimated with enough accuracy and the two sets of posterior estimates differed only in the third decimal place, signifying that the results are not sensitive to the choice of prior distributions for ξ_{vj}. In addition, the small values of the Monte Carlo standard errors of estimates suggested that the Markov chains with a run length of 10,000 and a burn-in period of 5,000 reached the stationary distribution. Further, note that the procedure recovered the latent structure accurately as well, as the posterior estimates of the correlation between the two distinct latent traits, displayed in the last row of Table 2, was close to the actual correlation in all the three situations. For this example where 2,000-by-16 data matrices were considered, each implementation took less than 25 minutes. The length of the chains may be increased to be as long as 50,000, which takes about 90-120 minutes for each execution.

Conclusion

This Fortran subroutine allows the user to choose between uniform and normal priors for the item parameters, α_v and γ_v. In addition, the user can modify the source code by assigning other values to μ_α, σ^2_α, and μ_γ, σ^2_γ to reflect different prior beliefs on their distributions. Convergence can be assessed by inspecting Monte Carlo standard errors, as well as by comparing the marginal posterior mean and standard deviation of each parameter computed for every 1,000 samples after the burn-ins. For the latter, identical values provide a rough indication of similar marginal posterior densities, which further indicates possible convergence of the Markov chain (Gelfand, Hills, Racine-Poon & Smith, 1990; Hoijtink & Molenaar, 1997).

Note that the algorithm adopts a correlation matrix in the prior distribution, $\boldsymbol{\theta}_i \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$, to solve the problem of model nonidentifiability (see e.g., Lee, 1995, for a description of the problem). Bafummi, Gelman, Park, and Kaplan (2005) provides an alternative solution to the problem.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171-187.

Brainerd, W. (2003). The importance of Fortran in the 21st century. *Journal of Modern Statistical Methods*, *2*, 14-15.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167-174.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (*2nd Ed.*). London: Chapman & Hall.

Table 1: Posterior Estimates and Monte Carlo Standard Errors of Estimates (MCSEs) for α_v with Uniform and Normal Priors

	$\rho = .2$		$\rho = .5$		$\rho = .8$	
	Uniform	Normal	Uniform	Normal	Uniform	Normal
Parameters	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)
α_1						
0.0966	0.0838 (.0013)	0.0828 (.0011)	0.0869 (.0007)	0.0846 (.0012)	0.0830 (.0013)	0.0847 (.0007)
0.0971	0.0675 (.0010)	0.0660 (.0013)	0.0731 (.0008)	0.0740 (.0010)	0.0657 (.0014)	0.0689 (.0012)
0.4589	0.4698 (.0035)	0.4704 (.0026)	0.4748 (.0028)	0.4707 (.0021)	0.4829 (.0021)	0.4797 (.0021)
0.9532	0.8556 (.0039)	0.8531 (.0069)	0.8804 (.0054)	0.8753 (.0058)	0.8937 (.0063)	0.8928 (.0045)
0.0771	0.0510 (.0009)	0.0502 (.0005)	0.0552 (.0013)	0.0550 (.0008)	0.0589 (.0007)	0.0577 (.0008)
0.4891	0.4900 (.0020)	0.4895 (.0024)	0.4855 (.0029)	0.4864 (.0012)	0.4659 (.0017)	0.4649 (.0017)
0.8599	1.0401 (.0185)	1.0348 (.0114)	1.0180 (.0080)	1.0120 (.0069)	0.9983 (.0057)	0.9930 (.0061)
0.9427	0.9381 (.0075)	0.9327 (.0024)	0.9477 (.0085)	0.9408 (.0088)	0.9628 (.0033)	0.9479 (.0075)
α_2						
0.2727	0.3013 (.0010)	0.2973 (.0026)	0.2654 (.0006)	0.2685 (.0014)	0.2348 (.0016)	0.2358 (.0013)
0.6532	0.7279 (.0051)	0.7251 (.0061)	0.6354 (.0028)	0.6346 (.0020)	0.7188 (.0042)	0.7142 (.0028)
0.1002	0.1231 (.0010)	0.1226 (.0014)	0.1528 (.0008)	0.1527 (.0012)	0.1088 (.0012)	0.1108 (.0018)
0.2339	0.0945 (.0014)	0.0965 (.0026)	0.1557 (.0021)	0.1535 (.0015)	0.1683 (.0020)	0.1670 (.0013)
0.9291	0.8554 (.0155)	0.8552 (.0131)	0.8145 (.0042)	0.8184 (.0071)	0.9208 (.0039)	0.9149 (.0061)
0.8618	0.8730 (.0128)	0.8575 (.0095)	0.9107 (.0060)	0.9001 (.0069)	0.9067 (.0034)	0.9055 (.0050)
0.0908	0.0543 (.0006)	0.0518 (.0016)	0.0556 (.0005)	0.0570 (.0007)	0.0463 (.0010)	0.0464 (.0007)
0.2083	0.2003 (.0006)	0.1967 (.0021)	0.2045 (.0016)	0.2035 (.0010)	0.2339 (.0013)	0.2351 (.0007)

Table 2: Posterior Estimates and Monte Carlo Standard Errors of Estimates (MCSEs) for γ_v and ρ with Uniform and Normal Priors

	$\rho = .2$		$\rho = .5$		$\rho = .8$	
	Uniform	Normal	Uniform	Normal	Uniform	Normal
Parameters	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)	Estimate (MCSE)
γ_1						
0.3629	0.3457 (.0007)	0.3447 (.0003)	0.3467 (.0010)	0.3448 (.0005)	0.3450 (.0004)	0.3452 (.0005)
-0.9010	-0.8881 (.0003)	-0.8875 (.0002)	-0.8891 (.0006)	-0.8885 (.0006)	-0.8875 (.0005)	-0.8865 (.0003)
-0.9339	-0.9288 (.0017)	-0.9277 (.0017)	-0.9288 (.0018)	-0.9270 (.0012)	-0.9317 (.0015)	-0.9310 (.0011)
-0.3978	-0.3976 (.0023)	-0.3983 (.0017)	-0.4035 (.0018)	-0.4012 (.0016)	-0.4059 (.0020)	-0.4062 (.0018)
0.3987	0.4077 (.0003)	0.4076 (.0008)	0.4085 (.0006)	0.4072 (.0006)	0.4073 (.0002)	0.4066 (.0007)
0.1654	0.1679 (.0003)	0.1681 (.0005)	0.1675 (.0009)	0.1666 (.0007)	0.1665 (.0010)	0.1669 (.0008)
-0.8108	-0.8302 (.0082)	-0.8294 (.0062)	-0.8232 (.0032)	-0.8186 (.0039)	-0.8122 (.0030)	-0.8091 (.0030)
-0.8012	-0.7091 (.0025)	-0.7064 (.0019)	-0.7145 (.0043)	-0.7102 (.0043)	-0.7186 (.0012)	-0.7140 (.0048)
γ_2						
0.2452	0.2902 (.0008)	0.2896 (.0007)	0.3122 (.0005)	0.3109 (.0002)	0.3037 (.0006)	0.3047 (.0005)
0.9792	1.0954 (.0031)	1.0941 (.0032)	1.0476 (.0015)	1.0461 (.0024)	1.1095 (.0021)	1.1045 (.0016)
-0.0190	-0.0216 (.0006)	-0.0212 (.0005)	-0.0058 (.0006)	-0.0068 (.0002)	-0.0200 (.0005)	-0.0196 (.0009)
0.8749	0.9549 (.0005)	0.9536 (.0006)	0.9624 (.0008)	0.9616 (.0009)	0.9568 (.0014)	0.9538 (.0005)
-0.3119	-0.2139 (.0026)	-0.2143 (.0013)	-0.2049 (.0019)	-0.2068 (.0011)	-0.2250 (.0005)	-0.2256 (.0011)
0.2005	0.2902 (.0025)	0.2888 (.0024)	0.2781 (.0021)	0.2735 (.0019)	0.2777 (.0012)	0.2750 (.0022)
0.4626	0.4658 (.0011)	0.4638 (.0004)	0.4514 (.0004)	0.4501 (.0002)	0.4550 (.0005)	0.4545 (.0012)
0.7184	0.7528 (.0008)	0.7510 (.0007)	0.7485 (.0007)	0.7462 (.0007)	0.7738 (.0003)	0.7723 (.0013)
ρ	0.1850 (.0022)	0.1853 (.0018)	0.5209 (.0031)	0.5213 (.0036)	0.7872 (.0037)	0.7942 (.0041)

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327-335.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, *85*, 315-331.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Boca Raton: Chapman & Hall/CRC.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using posterior predictive checks. *Psychometrika*, *62*, 171-189.

Lee, H. (1995). *Markov chain Monte Carlo methods for estimating multidimensional ability in item response analysis.* Ph.D. Dissertation, University of Missouri, Columbia, MO.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Ripley, B. D. (1987). *Stochastic simulation*. NY: Wiley.

Sheng, Y., & Headrick, T. C. (2007). An algorithm for implementing Gibbs sampling for 2PNO IRT models. *Journal of Modern Applied Statistical Methods*, *6*, 341-249.

Sheng, Y., & Wikle, C. K. (2007). Comparing multi-unidimensional and unidimensional IRT Models. *Educational & Psychological Measurement*, *67*, 899-919.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distribution by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

Appendix: Fortran Subroutine

```
SUBROUTINE GSMU2(Y, N, K, M, MN, L, BURNIN, BN, UNIF, ITEM, PER, RPER)
C*********************************************************************
C Y = the n-by-K binary item response data
C N = the number of subjects
C K = the test length (total number of items)
C M = the number of subtests
C MN = an array with numbers of items in the m subtests
C L = the number of iterations using Gibbs sampling
C BURNIN = the early number of iterations that are to be discarded
C BN  = the number of batches
C UNIF = a 0-1 indicator with 0 specifying normal priors for item
C     parameters and 1 specifying uniform priors for them
C ITEM = a K-by-4 matrix of posterior estimates and MCSEs for item
C     parameters
C PER = a n-by-2m matrix of posterior estimates and MCSEs for person
C      traits
C RPER = a (m*(m-1)/2)-by-2 matrix of posterior estimates and MCSEs
C     for the correlation(s) between person traits
C*********************************************************************
      INTEGER      N, K, MN(M), L, Y(N,K), IRANK, INDX(M), UNIF, COUNT,
      &            BURNIN, BSIZE, BN
      REAL   A(K), G(K), TH(N,M), AA(K,M), ZLP(N,K), LP, Z(N,K), PHAT(K),
      &      U, PVAR(M, M), SIGMA(M,M), RSIG(M,M), PVAR1(M,M), RTH(M),
```

```
      &      BA(K), PMEAN1(M), PMEAN(M), X(N,2), XX(2,2), IX(2,2), RMN(M),
      &      ZV(N,1), XZ(2,1), AMAT(2, 2), BZ(2,1), AMU, GMU, AVAR, GVAR,
      &      AGMU(2,1), AGVAR(2,2), AGSIG(2,2), BETA(2), C(M,M), CTH(M,N),
      &      D(M,M), AV(L,K), GV(L,K), RHO(M,M,L), THV(N,M,L), ITEM(K,4),
      &      PER(N,2*M), SUM0, SUM1, SUM2, SUM3,  M0, M1, M2, M3, TOT,
      &      TOT1, TOT2, TOT3, SS, SS1, SS2, SS3, RPER(M*(M-1)/2,2),
      &      PRODA, VAR(M,M)
      DOUBLE PRECISION  BB, TMP
C***********************************************************************
C Connect to external libraries for normal (RNNOR), uniform (RNUN), and
C multivariate normal (RNMVN) random number generator, inverse (DNORIN)
C and CDF (ANORDF, DNORDF) for the standard normal distribution, and
C Cholesky factorization (CHFAC) routines
C***********************************************************************
      EXTERNAL    RNNOR, RNUN, ANORDF, CHFAC, DNORDF, DNORIN, RNMVN
C***********************************************************************
C Set initial values for item parameters a(v), g(v), person ability
C theta, and the hyperparameter sigma, so a(v)=2, g(v)=-Φ^-1(Σ_i y_ij / n)√5
C for all k(v) items, theta(v)=0 for all n person traits, and sigma=I
C***********************************************************************
      PHAT = SUM(Y, DIM = 1)
      DO 10 I = 1, K
            A(I) = 2.0
            G(I) = -ANORIN(PHAT(I)/FLOAT(N))*SQRT(5.0)
   10 CONTINUE
      DO 15 I = 1, N
      DO 15 J = 1, M
            TH(I, J) = 0.0
   15 CONTINUE
      DO 20 I=1,M
      DO 20 J=1,M
            SIGMA(I, J) = 0.0
            SIGMA(I, I) = 1.0
   20 CONTINUE
      RMN = FLOAT(MN)
C***********************************************************************
C Start iteration
C***********************************************************************
      COUNT = 0
      DO 30 IT = 1, L
            COUNT = COUNT + 1
            DO 40 I = 1, K
            DO 40 J = 1, M
                  AA(I, J) = 0.0
   40       CONTINUE
            JJ = 0
            DO 50 I = 1, M
                  J = 1
                  DO WHILE (J .LE. MN(I))
                        JJ = JJ+1
                        AA(JJ, I) = A(JJ)
                        J = J + 1
                  END DO
   50       CONTINUE
```

```
C************************************************************************
C Update samples for Z from its normal posterior distributions
C************************************************************************
             ZLP = MATMUL(TH, TRANSPOSE(AA))
             DO 60 I = 1, N
             DO 60 J = 1, K
                   LP = ZLP(I, J) - G(J)
                   BB = ANORDF(0.0 - LP)
                   CALL RNUN(1, U)
                   TMP = BB*(1-Y(I, J)) + (1-BB)*Y(I, J))*U + BB*Y(I, J)
                   Z(I, J) = DNORIN(TMP) + LP
   60        CONTINUE
C************************************************************************
C Update samples for theta from their MVN posterior distributions
C************************************************************************
C************************************************************************
C Call the matrix inversion routine.
C Invert matrix SIGMA with the inverse stored in RSIG
C************************************************************************
            CALL MIGS(SIGMA, M, RSIG, INDX)
            PVAR1 = RSIG + MATMUL(TRANSPOSE(AA), AA)
C************************************************************************
C Call the matrix inversion routine to invert matrix PVAR1 with the
C inverse stored in PVAR
C************************************************************************
            CALL MIGS(PVAR1, M, PVAR, INDX)
            DO 70 I = 1, N
                  DO 80 J = 1, K
                        BA(J) = Z(I, J) + G(J)
   80             CONTINUE
                  PMEAN1 = MATMUL(TRANSPOSE(AA), BA)
                  PMEAN = MATMUL(PVAR, PMEAN1)
C************************************************************************
C Call the Cholesky factorization routine. Compute the Cholesky factorization
C of the symmetric definite matrix PVAR and store the C result in RSIG
C************************************************************************
                  CALL CHFAC (M, PVAR, M, 0.00001, IRANK, RSIG, M)
C************************************************************************
C Generate a random sample of theta(v) from MVN dist by calling RNMVN
C************************************************************************
                  CALL RNMVN (1, M, RSIG, M, RTH, 1)
                  DO 90 J = 1, M
                        TH(I, J) = RTH(J) + PMEAN(J)
                        THV(I, J, COUNT) = TH(I, J)
   90             CONTINUE
   70       CONTINUE
C************************************************************************
C Update samples for item parameters, a(v) and g(v) from their MVN
C posterior distributions
C************************************************************************
            JJ = 0
            DO 100 J = 1, M
                  DO 110 I = 1, N
                        X(I, 1) = TH(I, J)
                        X(I, 2) = -1
  110             CONTINUE
                  IF (UNIF == 0) THEN
```

```
C*************************************************************************
C Specify the prior means and variances for a(v) and g(v)
C*************************************************************************
                        AMU = 0.0
                        GMU = 0.0
                        AVAR = 1.0
                        GVAR = 1.0
C*************************************************************************
C Put them in vector or matrix format
C*************************************************************************
                        AGMU(1, 1) = AMU
                        AGMU(2, 1) = GMU
                        AGVAR(1, 1) = AVAR
                        AGVAR(2, 2) = GVAR
C*************************************************************************
C Call the matrix inversion routine to invert matrix AGVAR with the
C inverse stored in AGSIG
C*************************************************************************
                        CALL MIGS(AGVAR, 2, AGSIG, INDX)
                        XX = MATMUL(TRANSPOSE(X), X) + AGSIG
                  ELSE IF (UNIF == 1) THEN
                        XX = MATMUL(TRANSPOSE(X), X)
                  END IF
C*************************************************************************
C Call the matrix inversion routine to invert matrix XX with the
C inverse stored in IX
C*************************************************************************
                  CALL MIGS(XX, 2, IX, INDX)
C*************************************************************************
C Call the Cholesky factorization routine. Compute the Cholesky
C factorization of the symmetric definite matrix IX and store the
C result in AMAT
C*************************************************************************
                  CALL CHFAC (2, IX, 2, 0.00001, IRANK, AMAT, 2)
                  JM = 0
                  PRODA = 1.0
  130             JM = JM + 1
                  JJ = JJ + 1
                  DO 120 I = 1, N
                        ZV(I, 1) = Z(I, JJ)
  120             CONTINUE
                  IF (UNIF == 0) THEN
                        XZ = MATMUL(AGSIG, AGMU) + MATMUL(TRANSPOSE(X), ZV)
                  ELSE IF (UNIF == 1) THEN
                       XZ = MATMUL(TRANSPOSE(X), ZV)
                  END IF
                  BZ = MATMUL(IX, XZ)
                  A(JJ) = 0
                  DO WHILE (A(JJ) .LE. 0)
                        CALL RNMVN(1, 2, AMAT, 2, BETA, 1)
                        A(JJ) = BETA(1) + BZ(1, 1)
                        G(JJ) = BETA(2) + BZ(2, 1)
                  END DO
                  AV(COUNT, JJ) = A(JJ)
                  GV(COUNT, JJ) = G(JJ)
                  END IF
```

```
                  PRODA = PRODA*A(JJ)
                  IF (JM .LT. MN(J)) THEN
                        GOTO 130
                  END IF
                  DO 135 I = 1, M
                        C(I, J) = 0.0
  135             CONTINUE
                  C(J, J) = PRODA ** (1/RMN(J))
  100       CONTINUE
C*********************************************************************
C Update samples for the hyperparameter, SIGMA
C*********************************************************************
            CTH = MATMUL(C, TRANSPOSE(TH))
            D = MATMUL(CTH, TRANSPOSE(CTH))
C*********************************************************************
C Call the subroutine to generate the unconstrained covariance matrix
C VAR from the inverse Wishart distribution
C*********************************************************************
            CALL INVWISHRND(D, M, N, VAR)
            DO 140 I = 1, M
            DO 140 J = 1, M
                  SIGMA(I, J) = VAR(I, J)/SQRT(VAR(I, I))/SQRT(VAR(J, J))
                  RHO(I, J, COUNT) = SIGMA(I, J)
  140       CONTINUE
   30 CONTINUE
C*********************************************************************
C Calculate the batch means and se's for a(v), g(v), theta(v) and
C their correlations, and store them in ITEM, PER, and RPER
C*********************************************************************
      BSIZE = (L - BURNIN)/BN
      DO 150 J = 1, K
            COUNT = BURNIN
            TOT1 = 0.0
            TOT2 = 0.0
            SS1 = 0.0
            SS2 = 0.0
            DO 160 JJ = 1, BN
                  SUM1 = 0.0
                  SUM2 = 0.0
                  DO 170 I = 1, BSIZE
                        COUNT = COUNT + 1
                        SUM1 = SUM1 + AV(COUNT, J)
                        SUM2 = SUM2 + GV(COUNT, J)
  170             CONTINUE
                  M1 = SUM1/FLOAT(BSIZE)
                  M2 = SUM2/FLOAT(BSIZE)
                  TOT1 = TOT1 + M1
                  TOT2 = TOT2 + M2
                  SS1 = SS1 + M1*M1
                  SS2 = SS2 + M2*M2
  160       CONTINUE
            ITEM(J, 1) = TOT1/FLOAT(BN)
            ITEM(J, 2) = SQRT((SS1-(TOT1*TOT1/BN))/(BN-1))/SQRT(FLOAT(BN))
            ITEM(J, 3) = TOT2/BN
            ITEM(J, 4) = SQRT((SS2-(TOT2*TOT2/BN))/(BN-1))/SQRT(FLOAT(BN))
```

```
150 CONTINUE
    JJ = 0
    JK = 0
    DO 180 IM = 1, M
          JJ = JK + 1
          JK = JJ + 1
         DO 190 J = 1, N
                COUNT = BURNIN
                TOT3 = 0.0
                SS3 = 0.0
                DO 200 IB = 1, BN
                      SUM3 = 0.0
                      DO 210 I = 1, BSIZE
                            COUNT = COUNT + 1
                            SUM3 = SUM3 + THV(J, IM, COUNT)
210                   CONTINUE
                      M3 = SUM3/FLOAT(BSIZE)
                      TOT3 = TOT3 + M3
                      SS3 = SS3 + M3*M3
200             CONTINUE
          PER(J, JJ) = TOT3/FLOAT(BN)
          PER(J, JK) = SQRT((SS3-(TOT3*TOT3/BN))/(BN-1))/SQRT(FLOAT(BN))
190       CONTINUE
180 CONTINUE
    JK = 0
    DO 220 J = 1, M
    DO 220 IM = J + 1, M
          JK = JK + 1
          COUNT=BURNIN
          TOT = 0.0
          SS = 0.0
          DO 230 JJ = 1, BN
                SUM0 = 0.0
                DO 240 I = 1, BSIZE
                      COUNT = COUNT + 1
                      SUM0 = SUM0 + RHO(J, IM, COUNT)
240             CONTINUE
                M0 = SUM0/FLOAT(BSIZE)
                TOT = TOT + M0
                SS = SS + M0*M0
230       CONTINUE
          RPER(JK, 1) = TOT/FLOAT(BN)
          RPER(JK, 2) = SQRT((SS-(TOT*TOT/BN))/(BN-1))/SQRT(FLOAT(BN))
220 CONTINUE

    RETURN
    END
```

```
      SUBROUTINE INVWISHRND(S, P, V, IW)
C*********************************************************************
C S = p-by-p symmetric, positive definite 'scale' matrix
C P = order of the scale matrix
C V = 'degree of freedom parameter'
C                    (V must be an integer for this routine)
C IW = random matrix from the inverse Wishart distribution
C Note:
C     different sources use different parameterizations w.r.t. V.
C     this routine uses the model that
C     density (IW) is proportional to
C                  exp[-.5*trace(S*inv(IW))]/[det(IW)^((V+p+1)/2)]
C     With this density definition:
C                  mean(IW) = S/(V-p-1)
C*********************************************************************
      INTEGER  P, V, IRANK, INDX(P)
      REAL     S(P, P), IS(P, P), IW(P, P), W(P, P), Z(V, P), ZZ(P, P),
      &        A(P, P), AZ(P, P)

      DO 10 I = 1, V
      DO 10 J = 1, P
      CALL RNNOR (1, Z(I, J))
   10 CONTINUE
      ZZ = MATMUL(TRANSPOSE(Z), Z)
      CALL MIGS(S, P, IS, INDX)
      CALL CHFAC (P, IS, P, 0.00001, IRANK, A, P)
      AZ = MATMUL(TRANSPOSE(A), ZZ)
      W = MATMUL (AZ, A)
      CALL MIGS(W, P, IW, INDX)

      RETURN
      END
```

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 659-669

1538 – 9472/09/$95.00

JMASM29: Dominance Analysis of Independent Data (Fortran)

Du Feng
Texas Tech University

Norman Cliff
University of Southern California

A Fortran 77 program is provided for an ordinal dominance analysis of independent two-group comparisons. The program calculates the ordinal statistic, *d*, and statistical inferences about δ. The source codes and an executable file are available at http://www.depts.ttu.edu/hdfs/feng.php.

Key words: ordinal statistic, dominance analyses, independent *d*, Fortran.

Introduction

The frequently encountered location comparison problem in behavioral and psychological research is usually answered by the two-sample *t*-test, comparing means of the two groups, or the parallel one-way ANOVA. However, it has been argued that ordinal alternatives to mean comparisons, such as the dominance analysis δ (Agresti, 1984; Cliff, 1991, 1993; Hettmansperger, 1984; Randles & Wolfe, 1979), have advantages over the classical ones, because data in the social sciences are often ordinal in nature. In addition, ordinal methods are invariant under monotonic transformation, and can be more robust than the traditional normal-based statistics methods when the parametric assumptions are violated (Caruso & Cliff, 1997; Cliff, 1993; Long, Feng, & Cliff, 2003). This dominance analysis, δ, is summarized by the ordinal statistic *d* which compares the proportion of times a score from one group or under one condition is higher than a score from the other, to the proportion of times when the reverse is true. The *d* method not only tests the H_0: δ = 0, but also allows for determination of confidence interval (CI) bounds.

Fligner and Policello (1981) introduced a robust version of the Wilcoxon-Mann-Whitney test (Mann & Whitney, 1947) for comparing the medians of two independent continuous distributions, and tested behavior of *d*, using the sample estimate of its variance. Cliff (1993) suggested a modification of Fligner and Policello's (1981) procedure by deriving an unbiased sample estimate of the variance of *d* and setting a minimum allowable value for it in order to increase the efficiency of the estimate and to eliminate impossible values. Delaney and Vargha (2002) used modifications of the CI for δ with Welch-like *df*s, but these modifications did not take into account specific situations in which *d* with traditional CI performed poorly. Long, et al. (2003) proposed a further adjustment on the CI to account for boundary effects on the variance of *d* due to the negative correlation between σ_d^2 and δ. Simulation studies have shown that independent *d*, when compared to the *t*-test with Welch's adjusted *df* (Welch, 1937), behaves quite well in small and moderate samples under various normal and non-normal distributions in terms of Type I error rate, power, and coverage of the CI (Feng & Cliff, 2004).

Popular statistical software packages do not include ordinal dominance analyses. Thus, the purpose of this article is to provide a Fortan program that calculates the ordinal statistic, *d*, and statistical inferences about δ , for independent groups. The program also performs Welch's *t*-test on the same data for comparison.

Methodology

Independent *d* Analysis

The calculation of independent *d* involves comparison of each of the scores in one group to each of the scores in the other group . A

Du Feng is a Professor in the department of Human Development and Family Studies. Email: du.feng@ttu.edu. Norman Cliff is Professor Emeritus. Email: nrcliff5@q.com.

dominance variable d_{ij} is defined as: $d_{ij} = sign(x_i - x_j)$, where x_i represents any observation in the first group, x_j in the second. The d_{ij} simply represent the direction of differences between the x_i scores and the x_j scores: a score of +1 is assigned if $x_i > x_j$; a score of -1 is assigned if $x_i < x_j$; and a score of 0 is assigned if $x_i = x_j$. The d is an unbiased estimate of δ:

$$d = \Sigma\Sigma\, d_{ij} / \, n_1 n_2 \qquad (1)$$

whereas s_d^2, the unbiased sample estimate of σ_d^2, is obtained by

$$s_d^2 = \frac{n_1^2\Sigma(d_{i.} - d)^2 + n_2^2\Sigma(d_{.j} - d)^2 - \Sigma\Sigma(d_{ij} - d)^2}{n_1 n_2 (n_1 - 1)(n_2 - 1)} \qquad (2)$$

where $d_{i.}$ is

$$d_{i.} = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{n_1} \qquad (3)$$

and similarly for $d_{.j}$. To eliminate possible negative estimate of variance, $(1 - d^2)/(n_1 n_2 - 1)$ was used as the minimum allowable value for s_d^2. An asymmetric CI for δ was shown to improve the performance of d (Cliff, 1993; Feng & Cliff, 2004):

$$\delta = \frac{d - d^3 \pm Z_{\alpha/2} s_d \left(1 - 2d^2 + d^4 + Z_{\alpha/2}^2 s_d^2\right)^{1/2}}{1 - d^2 + Z_{\alpha/2}^2 s_d^2} \qquad (4)$$

where $Z_{\alpha/2}$ is the 1- α/2 normal deviate. When d is 1.0, s_d reduces to zero, the upper bound for the CI for δ is 1.0, and the lower bound is calculated by

$$\delta = \frac{(n_{min} - Z_{\alpha/2}^2)}{(n_{min} + Z_{\alpha/2}^2)} \qquad (5)$$

where n_{min} is the smaller of the two sample sizes. When d equals -1.0, the solution is the negative of (5).

The Fortran Program

The Fortran program for the independent groups d analysis applies the algorithm of the above Equations (1), (2), (3), (4), and (5). The program is interactive, supplying prompts at several points. Data can be either read from a file or input from the keyboard; if input from the keyboard, data will be stored in a file. In both cases, any number of experimental variables is possible, but an analysis is conducted on only one variable at a time. After input, data are sorted within each group.

The program calculates the statistical inferences about δ, generating d and its variance, as well as the components of variance of d. The outputs include a CI for δ and the significance of d (a z-score), testing the null hypothesis. The program also calculates the dominance variable d_{ij}, and a dominance matrix for the variables analyzed is generated as a part of the outputs when the data are no more than 75 cases. Otherwise, only the statistics and their components are included in the outputs. In order to compare the d method with the classical test methods, the program also performs the classical t statistic for independent groups with Welch's adjustment of degrees of freedom. Table 1 shows an example of the output file the program generated when the sample size is 25 for both groups.

Conclusion

The ordinal method d does not involve excessive elaboration and complicated statistical analyses. Its concepts can be easily understood by non-statisticians. However, popular statistical software packages such as SAS and SPSS do not allow for ordinal dominance analyses. This Fortran program (see the appendix for source codes) for independent groups d analysis is easy to implement. Its outputs provide descriptive information, not only the null hypothesis is tested, but also a CI is provided. In addition, a dominance matrix is produced as a useful visual aid to the test. A comparison of d with Welch's t. also is provided. Furthermore, if the users have access to the IMSL library, the current source codes can be easily adapted and used in Monte Carlo studies to evaluate the performance of d in terms of Type I error rate, power, and CI coverage.

Table 1: An Example of Independent d Analysis for Two Small Samples

Ordered Scores				Dominance Diagram
Alcoholic		Non-alcoholic		
Score	$d_{i.}$	Score	$d_{.j}$	
1	–1.00	3	.92	----------0----------
4	-.72	3	.92	+++0----------0----------
6	-.56	3	.92	+++++0----------0----------
7	-.52	4	.88	++++++----------0----------
7	-.52	5	.84	++++++----------0----------
14	-.24	6	.80	+++++++++0----------
14	-.24	12	.60	+++++++++0----------
18	.40	12	.60	++++++++++++++++000------
19	.52	13	.60	++++++++++++++++++------
20	.52	14	.52	++++++++++++++++++------
21	.52	15	.44	++++++++++++++++++------
24	.68	15	.44	++++++++++++++++++++----
25	.68	15	.44	++++++++++++++++++++----
26	.68	15	.44	++++++++++++++++++++----
26	.68	15	.44	++++++++++++++++++++----
26	.68	16	.44	++++++++++++++++++++----
27	.72	18	.40	++++++++++++++++++++0---
28	.84	18	.40	++++++++++++++++++++00-
28	.84	18	.40	++++++++++++++++++++00-
30	.92	23	.12	++++++++++++++++++++-
33	.92	23	.12	++++++++++++++++++++-
33	.92	27	-.32	++++++++++++++++++++-
44	1.00	28	-.44	++++++++++++++++++++
45	1.00	28	-.44	++++++++++++++++++++
50	1.00	43	-.76	++++++++++++++++++++

Inferences About δ	
d	.389
s_d	.154
.95 confidence interval	(.063, .640)
z for d	2.530
Components of s_d^2	
$s_{d_{i.}}^2$	.394
$s_{d_{.j}}^2$	.207
$s_{d_{ij}}^2$	.831
Mean Comparisons	
t for means	2.322
Welch's df for t	44.484

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. NY: Wiley.

Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, *57*, 637-654.

Cliff, N. (1991). Ordinal methods in the study of change. In Collins, L.M. & Horn, J. (Eds.), 34-46. *Best methods for the analysis of change*. Washington, D.C.: American Psychological Association.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494-509.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, *7*, 485-503.

Feng, D., & Cliff, N. (2004). Monte Carlo evaluation of ordinal *d* with improved confidence interval. *Journal of Modern Applied Statistical Methods*, *3*, 322-332.

Fligner, M. A., & Policello, G. E. II (1981). Robust rank procedure for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 76, 162-168.

Hettmansperger, T. P. (1984). *Statistical inferences based on ranks*. NY: Wiley.

Long, J. D., Feng, D., & Cliff, N. (2003). Ordinal analysis of behavioral data. In J. Schinka, W. Velicer, and I. B. Weiner (Eds.), *Comprehensive handbook of psychology, volume two: research methods in psychology*. NY: Wiley.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. NY: Wiley.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Appendix: Fortran Program

```
C*************************************************************************C
C This program computes independent groups d-statistics (Cliff, 1996; Long et al.,   C
C 2003; Feng & Cliff, 2004) and provides their standard errors, confidence intervals,C
C and tests of hypotheses. The program is interactive, supplying prompts at several  C
C points. It should be noted that before doing the analyses, you should have         C
C arranged your data in the specified format.                                        C
C Data can be either read from a file or input from the keyboard. If input from the  C
C keyboard, data will be stored in a file.  Data must be entered casewise, that is,  C
C all the data for one case or person, then all for the next, etc., and we need to   C
C know the number of cases and variables.  Group membership must be entered as       C
C variable.                                                                          C
C If data are in an external file, they must be cases by variables. That is, all the C
C scores for the first case or subject, all for the second, etc. In both cases,      C
C there could be any number of experimental variables, but you can do an analysis on C
C only one variable at a time.  We need to know the number of cases, and the number  C
C of variables for each case, including the grouping variable before running the     C
C program.                                                                           C
C If the data are no more than 75 cases, a dominance matrix for the variables        C
C analyzed will be printed as part of the output. Otherwise, just the statistics and C
C their components will be included in the output.                                   C
C The program is supplied as a professional courtesy. It can be used or copied for   C
C any academic or research purpose. However, it should not be copied for any         C
C commercial purpose.  We do not know of any errors, but do not guarantee it to be   C
C errors-free.  Please understand that it was written by amateur programmers, and is C
C not intended to be of commercial quality.                                          C
C*************************************************************************C
```

```
      INTEGER I,J,NV,NP,JQ,JC,JPLU,JG(2),NPER(2),GAP,IG,
     &          JORDER(2000,2),NDCOL(2000),NDROW(2000)
      REAL  YY,DB,SSROW,SSCOL,MINI,NUM,VARD,DEN,M1,M2,
     &        VARDROW,VARDCOL,VARDIJ,SD,UPPER,LOWER,SUM1,SUM2,MINN,
     &        SUMSQ1,SUMSQ2,VARDIFF,MDIFF,TEE,Y(2000,2),Z(2000,50)
      REAL  DEL,SQIJ,Q1,Q2,Q12
      CHARACTER*1 ANS, PLUS(3),DFILE*18,SPLU(70),SSPLU*70,
     &           STR*45, OUTFILE*8
      DATA PLUS(1),PLUS(2),PLUS(3)/'-','0','+'/
C*************************************************************************C
C Read data from a file, or input from the keyboard.                      C
C*************************************************************************C
      WRITE(*,101)
101   FORMAT('This is inddelta.f for computing d statistics.',
     & 3X,'It is copyright 1992, Norman Cliff. Comments and',
     & 1X,'suggestions are solicitted.')
      WRITE(*,102)
102   FORMAT('Type b to bypass instructions,any other letter to',
     & 1X,'see them.')
      READ(*,'(A1)') ANS
      IF((ANS.EQ.'B').OR.(ANS.EQ.'b')) GOTO 80
      WRITE(*,103)
103   FORMAT('Data can be either read form a file or input',
     & 1X,'from the keyboard. If it is in a file,it must be cases',
     & 1X,'by variabls, i.e., all the scores for the first case')
      WRITE(*,104)
104   FORMAT(' or subject, all for the second,etc. If it is not',
     & 1X,'arranged that way, type E for exit and go arrange it.')
      READ (*,'(A1)') ANS
      IF ((ANS.EQ.'E').OR.(ANS.EQ.'e')) GOTO 1500
80    WRITE(*,105)
105   FORMAT('Type f if it is in a file or k if you will enter',
     & 1X,'it from the keyboard.')
      READ(*,'(A1)')ANS
      IF((ANS.EQ.'K').OR.(ANS.EQ.'k')) THEN
        WRITE(*,111)
111     FORMAT('Data will be stored in a file. Give its full',
     & 1X,'name and extension.')
        READ(*,'(A18)') DFILE
        WRITE(*,112)
112     FORMAT('Data must be entered casewise, that is, all the',
     & 1X,'scores for one case or person, then all for the next,',1X,
     & 'etc.. And we need to know the number of cases and variables.')
        WRITE(*,113)
113     FORMAT('Group membership should be entered as a',
     & 1X,'variable.')
        WRITE(*,114)
114     FORMAT('Scores, or variables, within each case must be',
     & 1X,'separated by a comma.')
        WRITE(*,115)
115     FORMAT('No. of cases:')
        READ(*,'(I3)') NP
        WRITE(*,116)
116     FORMAT('No. of variables:')
        READ(*,'(I3)') NV
        OPEN(3,FILE=DFILE,STATUS='NEW')
        WRITE(*,117)
117     FORMAT(1X,'Enter the scores for each case, including',
     & 1X,'the grouping variable.')
        DO 1 I=1,NP
          WRITE(*,*) I
          DO 2 J=1,NV
             READ(*,*) Z(I,J)
```

```
2           CONTINUE
1         CONTINUE
         WRITE(*,118)
118      FORMAT(1X,'The scores will be printed out on the screen',
     & 1X,'for checking.')
         DO 3 I=1,NP
           WRITE(*,*) I
               WRITE(*,*) (Z(I,J),J=1,NV)
           WRITE(*,*)
3        CONTINUE
         WRITE(*,119)
119      FORMAT('If there are any corrections, type the row,',
     & 1X,'column, and the correct value. If not, type 0,0,0.')
276      READ(*,*) I,J,P
         IF(I.EQ.0) GOTO 281
         Z(I,J)=P
         WRITE(*,120)
120      FORMAT(1X,'More?  Type 0,0,0 , if not.')
         GOTO 276
281      DO 29 I=1,NP
            DO 30 J=1,NV
               WRITE(1,*) Z(I,J)
30           CONTINUE
29       CONTINUE
         CLOSE (3,STATUS='KEEP')
       ELSE
           IF((ANS.NE.'F').AND.(ANS.NE.'f')) THEN
                 GOTO 80
           ELSE
              WRITE(*,106)
106           FORMAT('Type name of file, including extention,',
     & 1X,'also path if not in this directory.')
              WRITE(*,107)
107           FORMAT('filename')
              READ(*,'(A18)') DFILE
              WRITE(*,108)
108           FORMAT('How many variables per case?')
              READ(*,'(I2)') NV
              WRITE(*,109)
109           FORMAT('How many cases?')
              READ(*,'(I3)') NP
              OPEN(4,FILE=DFILE,STATUS='OLD')
              DO 31 I=1,NP
                 READ(4,*) (Z(I,J), J=1,NV)
31            CONTINUE
              CLOSE(4,STATUS='KEEP')
           ENDIF
       ENDIF
282    WRITE(*,122)
122    FORMAT('Which variable no. is the grouping variable?')
       READ(*,'(I1)') JC
       WRITE(*,123)
123    FORMAT('Which variable no. is the experimental?')
       READ(*,'(I1)') JQ
       WRITE(*,124)
124    FORMAT('Which are two values of the grouping variable',1X,
     & 'designate the groups to be compared?(e.g.:1 and 2)')
       WRITE(*,125)
125    FORMAT(1X,'  First group: ')
       READ(*,'(I2)') JG(1)
       WRITE(*,126)
126    FORMAT(1X,'  Second group: ')
       READ(*,'(I2)') JG(2)
```

```
       NPER(1) = 1
       NPER(2) = 1
       WRITE(*,226)
226    FORMAT(1X,' Name of the output file is: ')
       READ(*,'(A9)') OUTFILE
       OPEN(8,FILE=OUTFILE)
C*************************************************************************C
C   Sort data.                                                            C
C*************************************************************************C
       DO 4  I=1,NP
        IF(Z(I,JC).EQ.JG(1)) THEN
          Y(NPER(1),1) = Z(I,JQ)
          JORDER(NPER(1),1) = NPER(1)
          NPER(1) = NPER(1)+1
        ELSE IF (Z(I,JC).EQ.JG(2)) THEN
          Y(NPER(2),2) = Z(I,JQ)
          JORDER(NPER(2),2) = NPER(2)
          NPER(2) = NPER(2)+1
        ELSE
        ENDIF
4      CONTINUE
        NPER(1)=NPER(1)-1
        NPER(2)=NPER(2)-1
        WRITE(*,127) NPER(1),NPER(2)
127     FORMAT(1X,2I4)
       DO 5  IG=1,2
         DO 6 K=4,1,-1
          GAP=2**K-K
            DO 7 I=GAP,NPER(IG)
             XX=Y(I,IG)
             YY=JORDER(I,IG)
             J=I-GAP
430          IF((J.LE.0).OR.(XX.GE.Y(J,IG))) GOTO 450
             Y(J+GAP,IG)=Y(J,IG)
             JORDER(J+GAP,IG)=JORDER(J,IG)
             J=J-GAP
             GOTO 430
450          Y(J+GAP,IG)=XX
             JORDER(J+GAP,IG)=YY
7           CONTINUE
6         CONTINUE
5       CONTINUE
C*************************************************************************C
C      Calculate dominance matrix (and print the matrix for small data set).   C
C*************************************************************************C
          SQIJ = 0.0
          DEL= 0.0
        WRITE(8,131)
131     FORMAT(1X,'This is an independent data analysis using',1X,
     & 'inddelta.f.')
        WRITE(8,*)
        WRITE(8,132) DFILE
132     FORMAT(1X,'The data are from ',A18)
        WRITE(8,*)
        WRITE(8,133) NV-1
133     FORMAT(1X,'There are ',I3,' experimental variable(s).')
        WRITE(8,*)
        WRITE(8,134) JC
134     FORMAT(1X,'The grouping variable is ',I3)
        WRITE(8,135) JQ
135     FORMAT(1X,'The experimental variable is ',I3)
        WRITE(8,*)
        DO 999 I = 1,NPER(1)
```

```
                NDROW(I) = 0
999     CONTINUE
        DO 998 I = 1,NPER(2)
                NDCOL(I) = 0
998          CONTINUE
         IF(NP.LE.75) THEN
           WRITE(8,137) JG(1),JG(2)
137        FORMAT(1X,'Dominance matrix for group',I3,' vs. group',I3)
           WRITE(8,*)
           WRITE(8,138) JG(1),JG(2)
138        FORMAT(1X,'A + INDICATES ',I3,' HIGHER THAN',I3)
           WRITE(8,*)
           DO 9 I=1,NPER(1)
                SSPLU = ' '
             DO 10 J=1,NPER(2)
                   IF(Y(I,1).GT.Y(J,2)) THEN
                      IWON=1
                   ELSE IF(Y(I,1).LT.Y(J,2)) THEN
                      IWON=-1
                   ELSE
                      IWON=0
                   ENDIF
                DEL = DEL +IWON
                SQIJ = SQIJ+IWON*IWON
                NDROW(I) = NDROW(I)+IWON
                NDCOL(J) = NDCOL(J)+IWON
                  JPLU = IWON + 2
                  SPLU(J) = PLUS(JPLU)
                  SSPLU = SSPLU(1:J)//SPLU(J)
10           CONTINUE
                WRITE(8,139) SSPLU
139             FORMAT(1X,A72)
9          CONTINUE
           WRITE(8,*)
           WRITE(8,*)
           WRITE(8,*)
         ELSE
           DO 11 I=1,NPER(1)
             DO 12 J=1,NPER(2)
                   IF(Y(I,1).GT.Y(J,2)) THEN
                      IWON=1
                   ELSE IF(Y(I,1).LT.Y(J,2)) THEN
                      IWON=-1
                   ELSE
                      IWON=0
                   ENDIF
                DEL = DEL +IWON
                SQIJ = SQIJ+IWON*IWON
                NDROW(I) = NDROW(I)+IWON
                NDCOL(J) = NDCOL(J)+IWON
12           CONTINUE
11          CONTINUE
         ENDIF
C*************************************************************************C
C     Calculate d and variance of d.                                      C
C*************************************************************************C
        DB = DEL/(NPER(1)*NPER(2))
        WRITE(8,*)
        WRITE(8,140)
140     FORMAT(1X,'***',2X,'d and its variance',2X,'***')
        WRITE(8,141) JG(1),JG(2),DB
141     FORMAT(1X,'d for ',I3,' vs. ',I3,27X,'  =  ',F6.3)
```

```
C*************************************************************************C
C       This part is for calculations of variance of d.                   C
C*************************************************************************C
         SSROW=0.0
         SSCOL=0.0
         DO 14 I=1,NPER(1)
         SSROW = SSROW + NDROW(I)**2
14       CONTINUE
         DO 15 I=1,NPER(2)
         SSCOL = SSCOL + NDCOL(I)**2
15       CONTINUE
         MINI=(SQIJ/(NPER(1)*NPER(2))-DB**2)
     &        /(NPER(1)*NPER(2)-1)
         NUM=SSROW-NPER(2)*DEL*DB + SSCOL - NPER(1)*DEL*DB
     &        -SQIJ + DEL*DB
         DEN = NPER(1)*NPER(2)*(NPER(1) - 1)*(NPER(2)-1)
         VARD = NUM/DEN
         IF (VARD.LE. MINI) THEN
            VARD = MINI
            WRITE(8,142)
142         FORMAT(1X,'variance = minimum.Interpret with caution.')
         ELSE
         ENDIF
         STR='variance for d'
         WRITE(8,143) STR,VARD
143      FORMAT(1X,A45,'  =  ',F7.4)
         VARDROW = (SSROW - NPER(2)*DEL*DB)
     &           /(NPER(2)**2*(NPER(1) - 1))
         VARDCOL = (SSCOL - NPER(1)*DEL*DB)
     &            /(NPER(1)**2*(NPER(2) - 1))
         VARDIJ = (SQIJ - DEL*DB)/(NPER(1)*NPER(2) - 1)
         WRITE(8,*)
         WRITE(8,144)
144      FORMAT(10X,'***  Components of the variance of d :  ***')
         STR='row di variance '
         WRITE(8,145) STR,VARDROW
145      FORMAT(1X,A45,'  =  ',F7.4)
         STR='column di variance '
         WRITE(8,146) STR,VARDCOL
146      FORMAT(1X,A45,'  =  ',F7.4)
         STR='variance of dij'
         WRITE(8,147) STR,VARDIJ
147      FORMAT(1X,A45,'  =  ',F7.4)
         SD = SQRT(VARD)
C*************************************************************************C
C   Calculate the asymmetric 95% confidence interval for delta,           C
C   with further agjustment on C.I. when d = 1.0 or d = -1.0.             C
C*************************************************************************C
         IF (NPER(1).LE.NPER(2)) THEN
            MINN = NPER(1)
         ELSE
            MINN = NPER(2)
         ENDIF
         IF (DB.EQ.1.0) THEN
            UPPER = 1.0
            LOWER = (MINN - 1.96**2)
     &              /(MINN + 1.96**2)
         ELSE IF (DB.EQ.(-1.0)) THEN
            LOWER = -1.0
            UPPER = -(MINN - 1.96**2)
     &              /(MINN + 1.96**2)
```

```
         ELSE
            UPPER = (DB-DB**3 + 1.96*SD*SQRT(DB**4 - 2*DB**2 + 1
     &+ 1.96*1.96*VARD)) / (1-DB**2 + 1.96*1.96*VARD)
             IF (UPPER.GT.1) UPPER = 1
            LOWER = (DB-DB**3 - 1.96*SD*SQRT(DB**4 - 2*DB**2 + 1
     &+ 1.96*1.96*VARD)) / (1-DB**2 + 1.96*1.96*VARD)
             IF (LOWER.LT.-1) LOWER = -1
         ENDIF
         WRITE(8,148)
148      FORMAT(10X,'**  Inference :  **')
         STR='approximate .95 Confidence limits for d '
         WRITE(8,149) STR
149      FORMAT(1X,A40)
         WRITE(8,*)
         IF(UPPER.GT.1) UPPER = 1
         IF(LOWER.LT.-1) LOWER = -1
         WRITE(8,150) LOWER,UPPER
150      FORMAT(20X,F6.3,' to ',F6.3)
         WRITE(8,*)
         STR='significance of d :'
         WRITE(8,151) STR,DB/SD
151      FORMAT(1X,A45,' z   =  ',F7.4)
         WRITE(8,*)
C*************************************************************************C
C        This short section computes the ordinary unpooled t-test         C
C        with Welch's adjusted df.                                        C
C*************************************************************************C
         SUM1 = 0.0
         SUM2 = 0.0
         SUMSQ1 = 0.0
         SUMSQ2 = 0.0
         DO 20 I = 1,NPER(1)
            SUM1 = SUM1 + Y(I,1)
20       CONTINUE
         M1 = SUM1/NPER(1)
         DO 21 I =1,NPER(1)
            SUMSQ1 = SUMSQ1 + (Y(I,1) - M1)**2
21       CONTINUE
         DO 22 I =1,NPER(2)
            SUM2 = SUM2 + Y(I,2)
22       CONTINUE
         M2 = SUM2/NPER(2)
         DO 23 I =1,NPER(2)
            SUMSQ2 = SUMSQ2 + (Y(I,2)-M2)**2
23       CONTINUE
         Q1 = SUMSQ1/(NPER(1)*(NPER(1)-1))
         Q2 = SUMSQ2/(NPER(2)*(NPER(2)-1))
         VARDIFF = Q1 + Q2
         MDIFF = M1 - M2
         TEE = MDIFF/SQRT(VARDIFF)
         WRITE(8,152)
152      FORMAT(6X,'***  Independent t-test with unpooled variance :',
     & 1X,' *** ')
         STR='mean difference'
         WRITE(8,153) STR,MDIFF
153      FORMAT(1X, A45,'  =  ',F7.4)
         STR='standard deviations:'
         WRITE(8,154) STR,SQRT((SUMSQ1/(NPER(1) - 1))),
     &               SQRT((SUMSQ2/(NPER(2)-1)))
154      FORMAT(1X,A47,'(1) ',F7.4,'  (2)   ',F7.4)
         STR='standard error of mean difference'
```

```
        WRITE(8,155) STR,SQRT(VARDIFF)
155     FORMAT(1X,A45,'  =  ',F7.4)
        Q12=(Q1+Q2)**2/(Q1**2/(NPER(1)-1)+Q2**2/(NPER(2)-1))
        WRITE(8,156) TEE,Q12
156     FORMAT(9X,'t  =  ',F8.4,9X,'adjusted df  =   ',F8.4)
        WRITE(8,*)
        WRITE(*,157)
157     FORMAT('Do you want the data to be printed on the',1X,
     &         'printer, y/n?')
        READ(*,'(A1)') ANS
        IF((ANS.EQ.'Y').OR.(ANS.EQ.'y')) THEN
            WRITE(8,*)
            WRITE(8,158)
158         FORMAT(10X,'***  Ordered data for this variable :  ***')
            WRITE(8,159)
159         FORMAT(1X,'ORDER',5X,'SUBJ.',5X,'SCORE',5X,'ROWDOM')
            WRITE(8,160) JG(1)
160         FORMAT(1X,'Group ',I3)
            DO 25 I=1,NPER(1)
                WRITE(8,161) I,JORDER(I,1),Y(I,1),NDROW(I)
161             FORMAT(1X,I5,5X,I5,5X,F6.3,5X,I3)
25          CONTINUE
            WRITE(8,162) JG(2)
162         FORMAT(1X,'Group ',I3)
            DO 26 I=1,NPER(2)
                WRITE(8,163) I,JORDER(I,2),Y(I,2),NDCOL(I)
163             FORMAT(1X,I5,5X,I5,5X,F6.3,5X,I3)
26          CONTINUE
        ELSE
        ENDIF
C*******************************************************************************C
        WRITE(8,*)
        WRITE(8,*)
        WRITE(8,*)
        WRITE(*,164)
164     FORMAT('Do you want to do another analysis, y or n?')
        READ(*,'(A1)') ANS
        IF (ANS.EQ.'Y'.OR.ANS.EQ.'y')  GOTO 282
1500    CLOSE(8,STATUS="KEEP")
        END
```